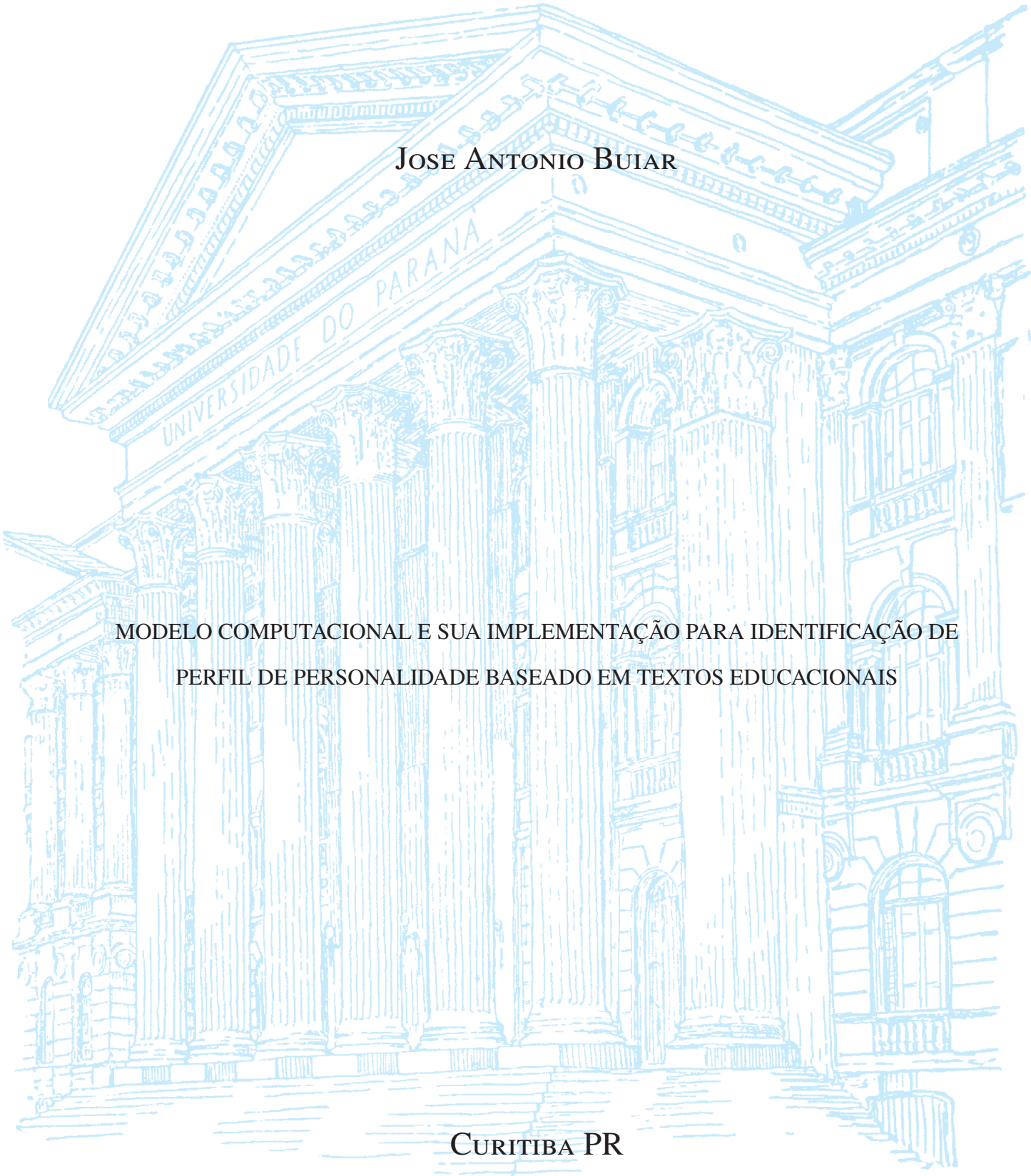
Universidade Federal do Paraná

Jose Antonio Buiar



MODELO COMPUTACIONAL E SUA IMPLEMENTAÇÃO PARA IDENTIFICAÇÃO DE PERFIL DE PERSONALIDADE BASEADO EM TEXTOS EDUCACIONAIS

Curitiba PR
2018

Jose Antonio Buiar

MODELO COMPUTACIONAL E SUA IMPLEMENTAÇÃO PARA IDENTIFICAÇÃO DE PERFIL DE PERSONALIDADE BASEADO EM TEXTOS EDUCACIONAIS

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Informática, no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Prof. Dr. Andrey Ricardo Pimentel.

Curitiba PR
2018

Catalogação na Fonte: Sistema de Bibliotecas, UFPR
Biblioteca de Ciência e Tecnologia

B932 Buiar, Jose Antonio
Modelo computacional e sua implementação para identificação de perfil de personalidade baseado em textos educacionais / Jose Antonio Buiar. – Curitiba, 2018.

Tese - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática, 2018.

Orientador: Dr. Andrey Ricardo Pimentel.

1. Processamento da linguagem natural (Computação). 2. Aprendizado de máquina. 3. Teste de personalidade (Criança). I. Universidade Federal do Paraná. II. Pimentel, Andrey Ricardo. III. Título.

CDD: 003.3

Bibliotecária: Lidiane do Prado Reis e Silva CRB-8/8579

MINISTÉRIO DA EDUCAÇÃO
SETOR SETOR DE CIENCIAS EXATAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO INFORMÁTICA

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **JOSE ANTONIO BUIAR** intitulada: **Modelo computacional e sua implementação para identificação de perfil de personalidade baseado em textos educacionais.**, após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua aprovação no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 14 de Setembro de 2018.

ANDREY RICARDO PIMENTEL
Presidente da Banca Examinadora

ROBERTO PEREIRA
Avaliador Interno

LUIZ EDUARDO SOARES DE OLIVEIRA
Avaliador Interno

MÁRCIA APARECIDA FERNANDES
Avaliador Externo

ROBINSON VIDA NORONHA
Avaliador Externo

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
Setor de Ciências Exatas
Coordenação de Pós-Graduação em Informática

Ao nosso eterno tutor inteligente, Alexandre Ibrahim Direne (in memoriam), por seu brilhante caminho no aperfeiçoamento da Informática na Educação e incansável apoio no início da condução desta pesquisa. O destino me conduziu a ser seu orientando, o destino fez com que eu fosse um dos seus últimos, o destino eternizou a sua obra.

Ao meu pai, Paulo Buiar (in memoriam), no centenário do seu nascimento, que mesmo sem completar o primeiro ano do ensino básico, não mediu esforços para oferecer as melhores condições de ensino a seus filhos. Tendo uma trajetória de vida exemplar, foi e continua sendo minha referência nos mais nobres valores que um ser humano possa almejar.

Agradecimentos

Aos meus pais pelo apoio incondicional durante toda minha vida e por terem proporcionado e incentivado meus estudos e participado na formação da minha personalidade.

A minha querida esposa Denise, aos meus filhos Analiz, Gabriel e Rafael e ao genro Tiago, por todo o apoio, incentivo e compreensão que foram fundamentais para a realização deste trabalho.

Ao meu orientador Andrey Ricardo Pimentel, por sua imensa generosidade em me adotar no meio do processo de doutoramento, bem como pelo incentivo e apoio durante a condução de seu nobre mister. Nossos saudáveis debates sobre as diversas áreas de conhecimento que envolveram este trabalho foram primordiais para o delineamento da solução apresentada.

Aos Professores Doutores convidados a compor a banca examinadora, Luiz Eduardo Soares de Oliveira, Márcia Aparecida Fernandes, Robinson Vida Noronha e Roberto Pereira, pela dedicação na avaliação do trabalho desenvolvido bem como por suas valiosas contribuições apresentadas durante a banca de defesa e posteriormente de forma escrita, que muito colaboraram para o aprimoramento do documento final.

Ao professor Luiz Eduardo Soares de Oliveira, que tanto me ajudou durante a realização do doutorado, sobremaneira nas atividades relacionadas a área de aprendizado de máquina, que nortearam a composição do modelo desenvolvido.

Ao meu grande amigo e orientador de mestrado, Robinson Vida Noronha, pesquisador de renome da área da Informática na Educação, pelo enorme incentivo a realização do doutorado e valiosos conselhos oferecidos durante a condução desta pesquisa.

Aos professores Lauro César Galvão, Iolanda Bueno de Camargo Cortelazzo e Marcus Vinicius Santos Kucharsk, docentes da Universidade Tecnológica Federal do Paraná (Curitiba), pelo grande apoio oferecido durante a realização dos experimentos realizados.

Aos meus colegas do Departamento Acadêmico de Informática da Universidade Tecnológica Federal do Paraná (Curitiba), pela colaboração e incentivo recebidos durante a realização desta pesquisa.

Aos colegas de curso, Carolina, Ernani e Zenaide, por todas as valiosas discussões, trabalhos conjuntos e disponibilidade para os revigorantes cafés, no Centro Politécnico e em outras partes do país, durante todos estes anos em que conduzimos nossas pesquisas.

À Universidade Tecnológica Federal do Paraná, por meio de seus gestores, por todo apoio recebido durante a realização deste doutorado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“ Quanto maior a incerteza, maior a lacuna entre o que você pode medir e o que importa, mais você deve ficar atento ao sobreajuste - ou seja, quanto mais você deva preferir a simplicidade”

Tom Griffiths

Resumo

A identificação do perfil de personalidade de alunos, levando em consideração as diferenças, colabora com os educadores no processo de encontrar situações de aprendizagem adequadas para cada aluno. Este processo pode ser realizado de forma intuitiva em pequenas turmas presenciais, mas apresenta-se como um grande desafio no cenário de grandes turmas em ambientes à distância. Uma das formas de identificar o perfil de personalidade é a utilização dos inventários de personalidade, nos quais os alunos respondem a uma série de perguntas que são posteriormente avaliadas, gerando os indicadores de perfil de personalidade de acordo com um modelo específico. Em contrapartida a esses métodos manuais de aplicação de inventários, tem-se desenvolvido métodos não intrusivos, baseados, por exemplo, na identificação das pistas de personalidade registradas pelos indivíduos nos textos por estes produzidos. Com a utilização de processos de aprendizado de máquina, as pistas identificadas nos textos podem ser comparadas às pistas identificadas em bases de dados de referência, nas quais um processo prévio de identificação manual foi realizado, inferindo-se assim o perfil de personalidade dos autores dos textos. Esta pesquisa apresenta um modelo, denominado IP3, que permite a realização da identificação automática do perfil de personalidade de alunos, de uma forma não intrusiva, tendo como referência somente o texto em português registrado por estes alunos em atividades educacionais. Este modelo é baseado em aprendizado de máquina, utilizando bases de aprendizado previamente rotuladas, modelos de representação do texto e técnicas de classificação. Como base de treinamento e referência para os testes dos classificadores, foram utilizadas as bases ESSAYS e *myPersonality*, bases estas utilizadas por diversas pesquisas na área de identificação de personalidade a partir do texto. Para a representação do texto foi utilizado o léxico LIWC, bem como a representação estatística nos modelos *n-gram* e *Word2Vec*. Também foram avaliadas as técnicas utilizadas para classificação de texto, sendo proposta a utilização da estratégia de combinação de classificadores. Com o objetivo de validar o modelo apresentado, foi realizado um experimento prático em um ambiente educacional. Os resultados apresentados demonstram a viabilidade da utilização do modelo IP3 para identificação do perfil de personalidade dos alunos baseado somente nos textos registrados em ambientes educacionais.

Palavras-chave: identificação de personalidade, classificação de texto, representação de texto, processamento de linguagem natural, aprendizado de máquina.

Abstract

The personality profile identification of students supports educators in the process of finding suitable learning conditions for each student while considering their differences. Although this process can be used in an intuitive way for small groups in classroom learning, it proves to be a significant challenge in the landscape of large distance learning groups. One way of identifying the personality profile is by using the personality inventory. By using this method, students answer a series of questions that are later evaluated, generating personality profile indicators according to a specific model. In contrast to that, we can find the use of non-intrusives methods. They are based on the identification of personality clues which can be derived from the text produced by individuals. With the use of machine learning processes, these clues are identified within the text and can be compared to the clues found in databases of reference, in which an inference of a personality profile has been identified, through a previous manual identification process. This research had the purpose of obtaining a model, named IP3, that allows the identification of students' profile in a non-intrusive way. It considered only text in portuguese produced by these students in their educational activities. To conduct this research, the author investigated text representation techniques that allowed to obtain clues about the writer. The methods used in this research were the LIWC lexicon as well as the statistic representation in the n-gram and Word2Vec models. Additionally, the classification and the classifiers combination specification techniques were also evaluated in the proposed model. As a training basis and reference for the classifiers' tests, ESSAYS and myPersonality databases have been used, which are commonly used by several researchers in the field of personality identification from text. To validate the model presented, a practical experiment was conducted in an educational environment. The presented results indicate the viability regarding the use of the IP3 student's personality profile identification model, based on the text produced by them during educational activities.

Keywords: personality recognition, text classification, text representation, natural language processing, machine learning.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

16PF	*Sixteen Personality Factor*
AMS	*Academic Motivation Scale*
ANN	*Artificial Neural Network*
ANEW	*Affective Norms for English Words*
APP	*Automatic Personality Perception*
APR	*Automatic Personality Recognition*
APS	*Automatic Personality Synthesis*
AVA	Ambiente Virtual de Aprendizagem
BIG FIVE	*Big Five Personality Traits*
BLR	*Bayesian Logistic Regression*
CBOW	*Continuous Bag of Words*
CNN	*Convolutional Neural Network*
CSV	*Comma-separated values*
DNN	*Deep Neural Network*
EaD	Educação a Distância
EAR	*Electronically Activated Recorder*
EPQ	*Eysenck Personality Questionnaire*
EPQR	*Eysenck Personality Questionnaire-Revised*
FC	*Fully-connected Neural Networks*
GNB	*Gaussian Naïve Bayes*
HTML	*Hypertext Markup Language*
IHC	Interação Humano Computador
ILS	*Index of Learning Styles*
JSON	*JavaScript Object Notation*
kNN	*k-Nearest Neighbors*
LIWC	*Linguistic Inquiry and Word Count*
LR	*Logistic Regression*
MBTI	*Myers-Briggs Type Indicator*
MEH	*Meaning Extraction Helper*
ML	*Machine Learning*
MLPC	*Multilayer Perceptron Classifier*
MRC	*Machine Usable Dictionary*
NB	*Naïve Bayes*

NEO-FFI	*NEO Five Factor Inventory*
NEO-PI-R	*NEO Personality Inventory Revised*
nGRAM	*n-gram Model*
NLTK	*Natural Language Toolkit*
NN	*Neural Network*
OCEAN	*Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism*
PLN	Processamento de Linguagem Natural
POS	*Part of Speech Tagger*
QP	*Quadratic Programming Problem*
RNN	*Recurrent Neural Network*
SCIKIT	*scikit-learn - Machine Learning in Python*
SMO	*Sequential minimal optimization*
SMS	*Short Message Service*
SNN	*Shallow Neural Network*
SPLICE	*Structured Programming for Linguistic Cue Extraction*
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency–inverse Document Frequency*
WE	*Word Embedding*
WEKA	*Waikato Environment for Knowledge Analysis*

Sumário

1 Introdução

O processo educacional, em decorrência das mudanças culturais que ocorreram na sociedade no último século, passou de um processo de difusão, onde a tarefa principal do professor era ser um propagador de ideias e conceitos, que deveriam ser assimilados pelos alunos por conta própria, para um processo mais participativo. No modelo inicialmente descrito, os alunos que não acompanhassem o processo simplesmente não eram educados adequadamente. Em um processo mais participativo, começam a ser relevantes as diferenças existentes entre os indivíduos, do ponto de vista do processo educacional. Conforme verificado por Perrenoud (2001), considerar as diferenças é encontrar situações de aprendizagem ótimas para cada aluno, buscando uma educação individualizada.

Um professor ao se deparar com sua nova turma, em uma primeira fase, até consegue identificar as diferenças relacionadas a classes sociais, religiosas, étnicas e de gênero de seus alunos, mas com o passar do tempo, o professor começa a perceber as diferenças comportamentais, suas atitudes e reações ao meio em que se encontram. Estes comportamentos podem ser entendidos como uma externalização da personalidade de cada indivíduo.

Observando estes comportamentos no dia a dia da sala de aula, o professor, mesmo que intuitivamente, realiza um processo de agrupamento dos alunos de acordo com estes padrões comportamentais. Por fim, acaba utilizando técnicas diferenciadas de ensino, mesmo que de modo tênue, usando uma entonação de voz diferenciada ou estimulando a participação maior de determinados alunos, para a condução do processo educacional. Quando o processo educacional ocorre com um número reduzido de educandos, em um modelo presencial, esta identificação das diferenças, mesmo com as naturais taxas de erro que irão ocorrer, é uma tarefa relativamente simples para o professor. Mas, quando envolve uma quantidade expressiva de alunos, quando a modalidade de educação é remota, ou realizada com suporte de Ambientes Virtuais de Aprendizagem (AVA), a identificação subjetiva por parte do professor fica altamente prejudicada, sendo esta identificação um desafio a ser vencido neste cenário.

Uma das formas de individualizar a diferença dos alunos é a aplicação de um método para identificar o perfil de personalidade. O método manual mais utilizado nesta área é a aplicação de questionários de avaliação de personalidade. Estes questionários são os mais válidos e confiáveis métodos atualmente disponíveis, para avaliação dos construtos de personalidade, são também medidas padronizadas e referenciadas por normas, que envolvem um conjunto padrão de perguntas, bem como um método de administração e a pontuação aplicada (Meyer et al., 2001). A utilização dos questionários de avaliação para alunos, na forma de perguntas de autoavaliação ou aplicados por especialistas, também apresenta os desafios práticos. A aplicação dos inventários tradicionais, que na maioria das vezes, são extensos e intrusivos, muitas vezes demandam tempo e outros recursos que nem sempre estão disponíveis para as aplicações do mundo real (Gosling et al., 2003). Nem sempre existem profissionais habilitados a aplicar os questionários aos alunos, bem como não pode ser assegurado que os alunos expressem naturalmente suas respostas em formulários *online* aplicado à distância.

Outra forma de investigar o perfil de personalidade dos alunos é buscar a identificação dos traços de personalidade através das pistas encontradas nas manifestações naturais dos indivíduos, como por exemplo, no texto redigido por estes, durante as atividades realizadas no decorrer do curso. A utilização das pistas encontradas no texto para a realização da identificação automática da personalidade, não especificamente no ambiente educacional, mas de forma mais generalizada, foi comprovada inicialmente pelos estudos de Argamon et al. (2005), Mairesse e Walker (2006) e Oberlander e Nowson (2006). A partir destas iniciativas, outras foram realizadas e publicadas nesta área, sendo que os métodos computacionais mais utilizados envolvem aprendizado de máquina. Os esforços de melhoria apresentados por estas iniciativas estão focados nos modelos de representação das pistas textuais, métodos de classificação e obtenção de bases robustas para treinamento.

O principal desafio para a utilização de aprendizagem de máquina, nos modelos investigados de identificação de personalidade a partir de texto, é a obtenção de bases de dados robustas, previamente classificadas para o treinamento do classificador. Em contrapartida, para um grande número de bases de dados de textos com classificação afetiva (positivo, neutro ou negativo), bases com classificação por meio de um modelo de identificação de personalidade, são escassas. A principal base disponível identificada, foi a desenvolvida por Pennebaker e King (1999), no idioma inglês, que serviu de referência para a realização de ensaios de diversos modelos apresentados na literatura. Outras iniciativas de classificação foram realizadas utilizando bases oriundas de redes sociais, coletando informações de usuários no *Facebook*, *Twitter* e *Youtube*, agregando informações adicionais ao texto, como informações dos perfis e das comunicações nestas redes. Este tipo de base de dados não está abrangido no escopo da presente pesquisa, que tem como foco apenas as informações textuais registradas pelos alunos em ambientes educacionais à distância. A obtenção de grandes bases textuais em si não é a dificuldade principal, mas sim, a classificação manual do perfil de personalidade dos autores dos textos, para permitir o treinamento dos classificadores. Neste aspecto, destaca-se a base obtida por Pennebaker, visto que a classificação foi realizada por especialistas em Psicologia.

Observa-se na investigação da literatura, que a utilização da identificação de personalidade a partir de textos, em ambientes educacionais no Brasil, ainda não tem sido amplamente utilizada. Não foi verificada também a disponibilidade de uma base de dados previamente classificada e publicamente disponível com textos em português. Todo este contexto foi a motivação inicial da presente pesquisa, desenvolver um modelo de identificação automática da personalidade de alunos, por meio dos textos produzidos em atividades educacionais, na língua portuguesa, que sirva de instrumento computacional para o auxílio do processo de ensino-aprendizagem. Estes meios poderiam ser estruturados em conteúdos programáticos, materiais didáticos e objetos educacionais, por exemplo, a serem aplicados de forma adaptável, de acordo com o perfil de personalidade de cada aluno.

Com isto temos o desenvolvimento da seguinte questão de pesquisa:

> Seria possível o desenvolvimento de um modelo computacional que permita, de forma automática e não intrusiva, a identificação do perfil de personalidade de alunos em um ambiente virtual de aprendizagem, baseado somente nos textos, em português, produzidos por estes nas atividades educacionais?

Esta questão central derivou algumas questões secundárias:

- Haveria um modelo de identificação que pudesse ser aplicado em ambientes educacionais, no Brasil, que permitisse a identificação do perfil de personalidade dos alunos, de forma automática e não intrusiva, baseado somente nos textos que os alunos redigiram em suas atividades educacionais?

- Qual técnica de identificação automática de personalidade poderia ser utilizada?
- Seria viável a especificação de um modelo de identificação automática que utilizasse uma base de dados em inglês para treinamento do classificador, em face da inexistência de bases em português, previamente classificadas?
- Seria este modelo facilmente adaptável a outras bases de classificação que forem sendo desenvolvidas ?
- Seria viável a realização de um estudo de caso para avaliar o modelo apresentado?

A busca por estas respostas é o fator motivador da presente pesquisa, que a partir de um processo investigativo, sobre as iniciativas já realizadas nesta área, apresenta um modelo que permite atingir os objetivos definidos nesta pesquisa.

1.1 Justificativa

A partir da revisão da literatura, apresentada nos Capítulos 3 e 4, não foram identificadas iniciativas abrangendo a identificação automática do perfil de personalidade baseado nos textos das atividades educacionais. Das pesquisas sobre identificação automática da personalidade que foram identificadas durante a condução desta pesquisa, 76% tiveram como base o idioma inglês. Além disto, foram utilizados textos obtidos em redes sociais ou outras fontes, e não baseadas em textos de ambientes educacionais. Esta constatação reforça a justificativa para a realização de uma pesquisa no escopo proposto pela tese aqui apresentada, cobrindo a lacuna verificada na investigação da literatura, referente a identificação do perfil de personalidade dos alunos, baseada em textos de ambientes educacionais em português, de forma não intrusiva e automática.

O desenvolvimento de um modelo computacional que permite a obtenção dos indicadores da personalidade dos alunos pode ser o ponto de partida para outras iniciativas complementares, que possam oferecer aos professores instrumentos que colaborem no processo de ensino-aprendizagem, sobremaneira no Ensino a Distância. Os artigos publicados em decorrência do desenvolvimento da presente pesquisa demonstram iniciativas realizadas na utilização da identificação automática da personalidade de alunos, baseado nos textos em português obtidos nas atividades educacionais, para a realização da inferência do estilo de aprendizado, no sequenciamento adaptativo de objetos de aprendizagem e na formação de grupos de colaboração.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo geral da presente pesquisa é propor um modelo de identificação automática do perfil de personalidade de alunos, que utilize a informação dos textos em português por estes registrados, nas atividades educacionais realizadas em um ambiente virtual de aprendizagem.

1.2.2 Objetivos Específicos

Como objetivos específicos desta pesquisa têm-se:

- avaliar os modelos de representação de textos para realização dos processos de classificação que permitam a identificação dos traços de personalidade a partir de textos educacionais;

- investigar as técnicas adequadas para classificação de texto com o objetivo de identificação dos perfis de personalidade dos alunos;
- verificar as bases de dados com perfil de personalidade disponíveis para utilização;
- especificar um modelo de identificação de personalidade a partir do texto dos alunos;
- realizar um experimento para validação e verificação do modelo proposto e verificação de suas contribuições e limitações.

1.3 Delimitações da Pesquisa

Esta pesquisa, fundamentada na Computação, tem característica multidisciplinar, envolvendo também as áreas de Educação e Psicologia, sendo importante delinear qual o escopo desenvolvido no presente trabalho, em cada uma destas áreas do conhecimento humano, conforme ilustrado na Figura 1.1. É importante salientar que esta pesquisa não tem a pretensão de trabalhar

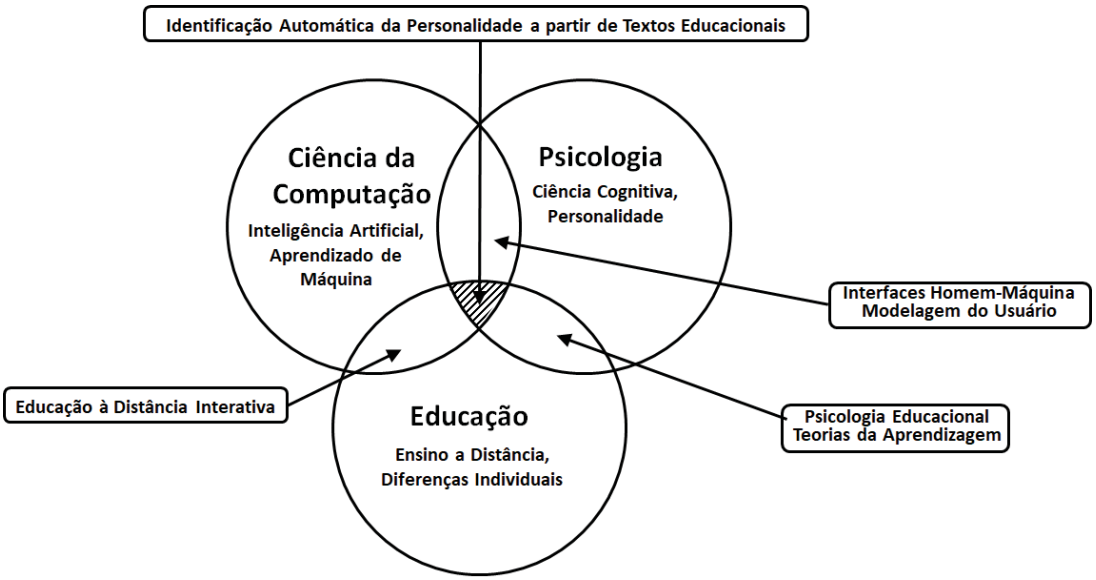


Figura 1.1: Interseção entre as Áreas do Conhecimento

os diversos modelos pedagógicos para o processo de ensino-aprendizagem. Assim, o axioma básico é de que cada individuo é um ser diferenciado no contexto do processo de ensino-aprendizagem, onde a hipótese concentra-se na utilização de um modelo computacional que permita reconhecer esta diferença, especificamente, pela identificação do perfil de personalidade dos alunos, onde isto poderia ser o ponto de partida para o desenvolvimento de meios didáticos que possam potencializar o processo de ensino-aprendizagem.

Considerando que de acordo com Matthews et al. (2003), os principais objetivos da Psicologia da Personalidade são distinguir as propriedades comportamentais internas de um indivíduo e investigar os relacionamentos causais entre eles, foi adotada a teoria dos traços de personalidade, por ser a modelagem que mais efetivamente prediz aspectos mensuráveis da vida de um individuo (Deary, 2009). Dentre os diversos modelos utilizados para a classificação da personalidade, utilizando a teoria dos traços de personalidade, foi escolhido como referência para a realização dos experimentos desta pesquisa, o modelo *Big Five Personality Traits* (BIG FIVE), sendo que o modelo computacional de identificação de personalidade apresentado na presente

pesquisa, pode ser facilmente adaptado a outro modelo de classificação da personalidade, caso necessário.

Neste trabalho não estão sendo discutidas as influências das questões de idioma, gênero e cultura no processo de identificação do perfil de personalidade. No modelo BIG FIVE, especificamente, por meio do estudo realizado por Schmitt et al. (2007) em em 56 países abrangendo diversos idiomas e culturas, foi comprovado que o inventário de identificação manual de personalidade, originalmente desenvolvido em inglês, pode ser adaptado a outros idiomas e culturas, apresentado altos níveis de confiabilidade interna.

O conceito de textos educacionais, como referência a fonte de informações utilizada pelo modelo apresentado para a realização da identificação dos traços da personalidade individual dos alunos, abrange os textos gerados pelos alunos em resposta a atividades propostas pelos professores ou tutores no Ambiente Virtual de Aprendizagem, e não abrange os textos considerados como material didático, como material de leitura, apostila e outros similares.

1.4 Estrutura do Trabalho

O presente trabalho está organizado na forma de capítulos. O Capítulo 2, com o objetivo de propiciar um aprofundamento crítico sobre o tema da pesquisa, apresenta uma breve revisão sobre os fundamentos teóricos envolvendo a Computação e a Personalidade, as técnicas de representação dos traços de personalidade presentes no texto e os modelos computacionais de classificação em aprendizado de máquina. Na sequência, o Capítulo 3 apresenta uma revisão da literatura abrangendo os estudos relacionando a Personalidade e Educação, seguido do Capítulo 4,onde é realizada uma revisão dos estudos na área de Identificação de Perfil de Personalidade baseado em Texto. O Capítulo 5 apresenta o modelo proposto pela pesquisa. Os experimentos realizados, os resultados obtidos e a validação do modelo estão presentes no Capitulo 6. No Capítulo 7 são realizadas as considerações finais e sugestões de trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo são apresentados os aspectos relacionados à identificação automática da personalidade baseada em texto. O conceito de Personalidade, suas características, modelos de representação e formas de avaliação são descritos na Seção 2.1. Na sequência, a Computação da Personalidade e seus diferentes aspectos são abordados. O relacionamento da Personalidade com a Educação é verificado na seção seguinte, enfatizando os aspecto motivacionais e cenários de aplicação. Uma seção descrevendo um modelo para Reconhecimento Automático da Personalidade a partir do Texto é então apresentado. Na próxima seção são apresentadas as Bases de Dados mais utilizadas na validação dos processos de identificação de personalidade a partir do texto. As técnicas de representação de texto são abordadas na seção seguinte, e na continuidade, são descritas as técnicas de classificação e seus conceitos relacionados. O capítulo é encerrado com as considerações sobre a fundamentação teórica. A importância destes conceitos para a presente pesquisa é extrair princípios e abordagens de sucesso, utilizados nos processos de identificação da personalidade a partir do texto, com o objetivo de buscar os fundamentos de soluções para o problema apresentado, que contribuam para elaboração do modelo proposto.

2.1 Personalidade

O presente estudo abrangerá as relações da Personalidade com a Computação. Esta seção oferece uma visão histórica sobre a descoberta da personalidade humana, os seus modelos e formas de avaliação. A Psicologia pode ser descrita como a área da ciência que estuda o comportamento e a mente, incluindo os fenômenos conscientes e inconscientes (Hockenbury e Hockenbury, 2010). Tem suas origens nas antigas civilizações representando uma abrangente e importante área do conhecimento humano. Suas principais vertentes são a biológica, comportamental, cognitiva, social, psicanalítica e teorias humanísticas existenciais. Como temas de estudo compreende a personalidade, o pensamento inconsciente, a motivação, o desenvolvimento psicológico e a genética comportamental.

A Psicologia da Personalidade realiza o estudo dos padrões de comportamento nos indivíduos, por meio da emoção. As teorias da personalidade estão ligadas a diversas escolas de pensamento e apresentam classificações e formas de mensuração distintas. O estudo das diferenças individuais dentro da Psicologia da Personalidade é um tema que também tem despertado interesse por parte dos profissionais de Computação, como por exemplo na área de IHC (Isbister e Nass, 2000; Karsvall, 2002; Saadé et al., 2006; Shami et al., 2008; Nov et al., 2013).

A personalidade pode ser descrita como um conjunto de diferenças individuais que são afetadas pelo desenvolvimento de um indivíduo: valores, atitudes, memórias pessoais, relacionamentos sociais, hábitos e habilidades (McAdams e Olson, 2010; Michel et al., 2004). Na Grécia antiga, Hipócrates (460 a.C. - 377 a.C.) classificava as pessoas em quatro tipos, de acordo com a presença de certas substâncias no organismo, as quais denominou de “humores”. Cláudio Galeno

(130 d.C. - 210 d.C.) expandiu a teoria de humorismo para teoria da personalidade, indicando que havia uma relação direta entre os níveis de humores do corpo, com as inclinações emocionais e comportamentais, os temperamentos, definindo-os em quatro tipos: sanguíneo; colérico; melancólico e fleumático. O filósofo grego Teofasto (372 a.C. — 287 a.C.) observou e descreveu as diferenças individuais , organizando-as com a descrição taxonômica de "Características". As características de Teofasto são frequentemente utilizadas para ilustrar e simbolizar a lacuna de coerência da descrição de traços de personalidade.

A Teoria Psicanalítica da Personalidade, desenvolvida por Sigmund Freud, defende que toda ação é movida por forças internas, que estão diretamente ligadas ao prazer, ou seja, para ele o desenvolvimento da personalidade é regido pela libido. Esta teoria indica que a personalidade é desenvolvida no indivíduo quando criança. A fim de explicar a sua teoria, Freud subdividiu a estrutura da personalidade em três sistemas: o "id", o "ego" e o "superego". A subdivisão do ponto de vista de Freud, explica os processos psicológicos trabalhando juntos, funcionando como um todo na personalidade, onde o "id" desempenha o fator biológico, o "ego" o psicológico e "superego" o social (Schultz e Schultz, 2016).

Na Teoria da Personalidade Junguiana (Jung, 2014), idealizada por Carl Gustav Jung, a personalidade, ou psique, é formada por sistemas isolados que atuam de forma dinâmica uns sobre os outros, não concordando com a visão de Freud sobre os fatores de motivação da personalidade. Na visão junguiana, existem quatro funções psicológicas básicas (sentir, pensar, perceber e intuir) e dois tipos de caráter (introvertido e extrovertido).

Desde que Allport (1937) e Murray (1938) imaginaram a psicologia da personalidade como o estudo científico da individualidade psicológica, psicologistas da personalidade, concentraram suas pesquisas nas mais importantes diferenças no funcionamento social e emocional que distinguem uma pessoa das outras. Cada vida humana é a variação de um projeto evolucionário geral, desenvolvido sobre o tempo e a cultura (McAdams e Pals, 2006). Allport (1937) encontrou mais de 50 significados distintos nas definições de personalidade pesquisadas. Em termos gerais pode-se dizer que:

> "*Personalidade é um conjunto de características individuais de padrões de pensamento, emoção e comportamento, junto com os mecanismos psicológicos, ocultos ou não, por detrás destes padrões (Funder, 1997)* ".

Os principais objetivos da psicologia da personalidade são: distinguir as propriedades comportamentais internas de um indivíduo e investigar os relacionamentos causais entre eles (Matthews et al., 2003). Uma das principais abordagens utilizadas para o estudo da personalidade humana é a teoria dos traços de personalidade. Esta teoria está baseada na mensuração de traços que podem ser definidos como padrões habituais de comportamento, pensamento e emoção (Kassin, 2003). O conceito de traços de personalidade é a modelagem que mais efetivamente prediz aspectos mensuráveis da vida dos indivíduos, largamente reconhecida como uma dos maiores realizações da psicologia (Deary, 2009). Segundo verificado por Matthews et al. (2003), esta teoria esta fundamentada em dois princípios:

- são estáveis ao longo do tempo e
- influenciam diretamente o comportamento humano.

Em algumas teorias, os traços são algo que um indivíduo possui ou não possui, enquanto em outras, os traços são dimensões tais como, introversão e extroversão, em que cada indivíduo recebe uma medida indicando onde está posicionado dentro deste espectro específico de variação.

Os traços de personalidade são características que não podem ser medidas com precisão, mas se um traço de personalidade for relevante, causando diferenças individuais significativas, este será notado (Goldberg, 1981).

Cattell et al. (1970) identificaram 16 fontes de traços de personalidade e definiram o Questionário dos Dezesseis Fatores de Personalidade (16PF). Este questionário tem sido utilizado para realizar levantamentos dos traços de personalidade dos indivíduos. A quinta edição deste questionário, mais atual, é datada de 1993. A quarta edição teve cinco variantes publicadas entre 1967 e 1969. As três edições iniciais são datadas de 1949, 1956 e 1962. As dimensões bipolares de Cattell estão apresentadas na Tabela 2.1.

Tabela 2.1: Dimensões Bipolares de Cattel

1	Expressividade emocional (alta-baixa)
2	Inteligência (alta-baixa)
3	Estabilidade (força do Eu-fraqueza do Eu)
4	Dominância (dominância-submissão)
5	Impulsividade (urgência/impulsividade-não-urgente)
6	Conformidade grupal (superego forte-superego fraco)
7	Atrevimento (atrevimento/timidez)
8	Sensibilidade (sensibilidade/dureza)
9	Desconfiança (confiança/desconfiança)
10	Imaginação (pragmatismo/imaginação)
11	Astúcia (sutileza/ingenuidade)
12	Culpabilidade (consciência-impassibilidade)
13	Rebeldia (radicalismo-conservadorismo)
14	Autossuficiência (autossuficiência/dependência)
15	Autocontrole (autoestima/indiferença)
16	Tensão (tensão-tranquilidade)

Em uma outra vertente mais simplificada de classificação da personalidade, o modelo de Três Fatores de Eysenck (Eysenck, 1947) propõe que o núcleo da personalidade consiste das três trilhas:

1. Introversão e Extroversão;
2. Neuroticismo e Estabilidade;
3. Psicoticismo.

Como instrumento de mensuração do modelo de Eysenck, foi desenvolvido o *Eysenck Personality Questionnaire* (EPQ), detalhado na Seção 2.1.2. Uma respeitável referência nesta área é a teoria dos tipos de Myers-Briggs (Myers et al., 1985) (MBTI), baseada nas teorias de Jung, está baseada em quatro dimensões que permitem categorizar a personalidade de um indivíduo. Estas dimensões bipolares estão ilustradas na Tabela 2.2.

Com posições próprias, diversos autores indicaram preferências por modelos que contêm um diferente número de fatores, desde Um, Dois, Cinco, Seis até Sete fatores, conforme ilustrado na Figura 2.1, com suas terminologias originais.

O modelo mais aceito e difundido nesta área é o *Big Five Personality Traits* (BIG FIVE), que tem sido amplamente utilizado em estudos da personalidade humana e na identificação dos

Tabela 2.2: Dimensões de Myers-Briggs

E	Extrovertido	x	Introvertido	I
S	Sensorial	x	Intuitivo	N
T	Racionalista	x	Sentimental	F
J	Julgador	x	Perceptivo	P

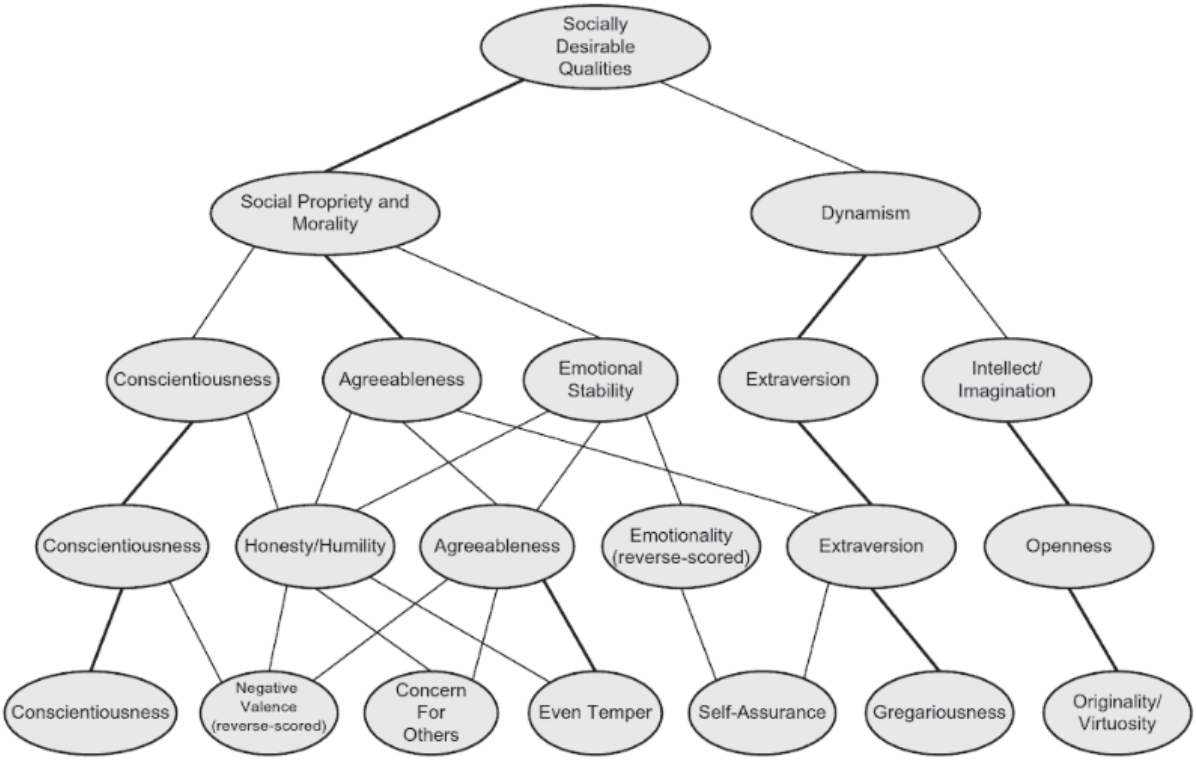


Figura 2.1: Relação entre as Estruturas de Um, Dois, Cinco, Seis e Sete Fatores

Fonte: Boyle et al. (2008)

indivíduos em função dos seus traços de personalidade. Este modelo está detalhado na Seção 2.1.1.

Os modelos de traços de personalidade também são amplamente aceitos na comunidade computacional. De acordo com o levantamento sobre Computação da Personalidade realizado por Vinciarelli e Mohammadi (2014), todos os 81 trabalhos pesquisados utilizaram estas teorias, com grande domínio do modelo BIG FIVE. Além de apresentarem o modelo de traços de personalidade como sendo o modelo dominante na Psicologia da Personalidade, os autores desta pesquisa indicam que este modelo representa a personalidade em termos de valores numéricos, uma forma particularmente adequada para processos computacionais.

2.1.1 Modelo BIG FIVE

O modelo BIG FIVE é produto de muitas décadas de pesquisa analítica centrada nas características da personalidade humana. Foi originalmente concebido por Galton (1949) tendo suas raízes nas hipóteses léxicas para identificação de traços de personalidade. A característica da abordagem léxica do modelo significa que os descritores de personalidade serão encontrados

nas evidências da linguagem natural (John et al., 1988; De Raad, 2000; Saucier e Goldberg, 2001; Saucier e Srivastava, 2015).

Conforme verificado por Goldberg (1990), as cinco dimensões da personalidade podem ter sido a base para diversas teorias de personalidade da época, como as propostas de Cattell (1957), Norman (1963), Eysenck (1970) e Guilford (1975). Na sequência, investigações empíricas demonstraram uma forte ocorrência de cinco domínios de personalidade, mas com nomenclatura ligeiramente diferente (Goldberg, 1990; Digman, 2002). Estas dimensões, também chamadas de fatores, estão ilustradas na Tabela 2.3.

Tabela 2.3: Fatores de Personalidade do Modelo BIG FIVE

Fator	**Descrição**
Openness to Experience	"abertura" é o interesse pela arte, emoção, aventura, ideias fora do comum, imaginação, curiosidade, e variedade de experiências
Conscientiouness	"conscienciosidade", ou meticulosidade, é a tendência para mostrar autodisciplina, orientação para os deveres e para atingir os objetivos
Extraversion	"extroversão" é caracterizada por emoções positivas e pela tendência para procurar estimulação e a companhia dos outros
Agreeableness	"amabilidade", ou socialização, é a tendência para ser compassivo e cooperante em vez de suspeitoso e antagonista, face aos outros
Neuroticism	"neuroticismo" é a tendência para experimentar emoções negativas, como raiva, ansiedade ou depressão

Fonte: Matthews et al. (2003)

Na sequência deste trabalho, estes fatores serão referenciados individualmente pelas palavras *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness* e *Neuroticism*, e de forma conjunta, pelo acrônimo OCEAN, que é formado pela letra inicial de cada um destes fatores.

$$OCEAN = [Openness,\ Conscientiousness,\ Extraversion,\ Agreeableness,\ Neuroticism] \quad (2.1)$$

Cada um destes fatores pode representar o perfil da personalidade de um indivíduo em relação a esta faceta específica. Tendo como exemplo a faceta de *Neuroticism*, no qual seu oposto é a estabilidade emocional, pessoas com escore elevado nesta faceta tendem a ter experiências tais como sentimentos negativos e instabilidade emocional, constrangimento, culpa, pessimismo e baixa autoestima (Zhang, 2003).

A utilização dos dados obtidos pelo questionário NEO-PI-R[1], aplicados em diversos países, incluindo o Brasil, por 78 membros colaboradores internacionais em cerca de 12 mil voluntários, verificou a validade da utilização deste questionário para identificação do BIG FIVE em um público heterogêneo abrangendo mais de 50 culturas diferentes (McCrae e Terracciano, 2005). Em outra iniciativa conduzida por Schmitt et al. (2007), foi realizado um levantamento sobre a distribuição geográfica do BIG FIVE. Esta pesquisa foi realizada em 29 idiomas, cobrindo 56 países dos cinco continentes, incluindo também o Brasil. Foram envolvidos mais de 100 cientistas procurando abranger as diversas culturas de cada continente. Estes estudos concluíram que a estrutura do inventário desenvolvido para o idioma inglês pode ser altamente aplicável

[1]*NEO Personality Inventory - revised* (Costa e McCrae, 1992), descrito na Seção 2.1.2.

na maioria das regiões culturais do mundo, indicando altos níveis de confiabilidade interna em todas as culturas.

O estudo dos traços de personalidade está intimamente ligado ao processo de levantamento e medição necessários para identificar as dimensões básicas da personalidade. O pesquisador possui tipicamente algumas hipóteses sobre o número e a natureza das principais dimensões e desenvolve um questionário para realizar a avaliação. Trabalhos posteriores investigam a utilidade e validade do questionário e os adaptam em função dos resultados obtidos (Matthews et al., 2003).

2.1.2 Avaliação da Personalidade

A utilização de questionários investigativos e questionários de autoavaliação tem dominado o campo de avaliação da personalidade. Outros métodos consistem em estratégias multimodais de avaliação para garantir convergência de resultados relacionados ao levantamento do perfil de personalidade, assim como diagnósticos de desordem cognitiva e afetiva, histórico do caso e outros dados como entrevistas, observações informações ambientais e todo o arcabouço de informações informais que possam ser agregado a testes padrões e questionários. Entretanto, conforme um trabalho publicado por um grupo de renomados pesquisadores da área:

> "*As medidas padronizadas e referenciadas por normas (conjunto de questões, métodos de administração, pontuação), ou seja, questionários de avaliação, são os mais válidos e confiáveis métodos atualmente disponíveis, para avaliar os construtos de personalidade (Meyer et al., 2001)*".

O *Eysenck Personality Questionnaire* (EPQ) foi originalmente desenvolvido por Eysenck e Eysenck (1976), sendo composto por 90 itens que têm como objetivo avaliar os traços de personalidade de um indivíduo. Os autores estimaram que a personalidade poderia ser avaliada com base em duas dimensões biologicamente independentes do temperamento, a dimensão E e a dimensão N. A dimensão E está relacionada ao conceito de Extroversão e Introversão. Cerca de 16% da população tende a estar na faixa da extroversão, 68% na faixa intermediária e os restantes 16% na faixa da introversão (Bartol e Bartol, 2014). A dimensão N está ligada aos conceitos de Neuroticismo e Estabilidade. O Neuroticismo está baseado nos níveis de ativação do sistema nervoso simpático, em reação a algum tipo de perigo ou ameaça. Os indivíduos que possuem um nível mais baixo de disparo deste gatilho, estão mais próximos da faixa de neuroticismo, ao passo que as pessoas que possuem um nível maior de ativação e maior controle emocional, estão mais próximos do nível de estabilidade (Eysenck e Eysenck, 1976). Posteriormente uma terceira dimensão P foi adicionada, sendo associada aos conceitos de Psicoticismo e Socialização. O conceito de psicoticismo está relacionado com a probabilidade de ocorrência de um episódio psicótico e também com a agressividade. Com a introdução de uma quarta dimensão L, associada com a escala da mentira (*Lie*), foi especificada uma versão revisada com 36 itens, da escala psicótica, conhecida como *Eysenck Personality Questionnaire-Revised* (EPQ-R) (Eysenck et al., 1985).

O trabalho realizado por Costa e McCrae (1985, 1992) pode ser considerado um dos mais relevantes na área de identificação de perfil de personalidade. De acordo com Taylor e MacDonald (1999), o *NEO Personality Inventory* (NEO-PI) (Costa e McCrae, 1985) não somente demonstra grandes propriedades psicométricas, como também consegue acomodar construtores já endereçados pelas métricas existentes de traços de personalidade. O termo NEO é um acrônimo formado com as iniciais das três fatores inicialmente incluídos no estudo: *Neuroticism*, *Extraversion* e *Openness to Experience*. Inicialmente, Costa e McCrae (1985) incluíram escalas

para medição de seis facetas conceitualmente derivadas dos três fatores NEO, mas não incluíram as facetas dos recentes fatores *Agreeableness* e *Conscientiouness*. Com a publicação do *NEO Personality Inventory - revised* - NEO-PI-R, Costa e McCrae (1992) ampliaram o questionário para abranger 240 itens, incluindo estes dois novos fatores. Ao contrário da maioria dos estudos léxicos da época, que eram baseados em amostras de estudantes de nível médio, o NEO-PI-R foi desenvolvido com amostras de estudantes de meia idade e adultos. Estas escalas demonstraram substancial consistência interna, estabilidade temporal, bem como convergência e validade discriminantes (McCrae e Costa, 2003; Costa e McCrae, 1992).

Este questionário é baseado em respostas para cada uma das questão e foram feitas de acordo com os cinco pontos da escala Likert (Allen e Seaman, 2007), indo de "Discordo Totalmente" até "Concordo Totalmente". Cada trilha é composta por seis sub-trilhas, cada uma associada a oito questões de avaliação. Um tempo estimado de 45 minutos é previsto para um indivíduo completar este questionário de autoavaliação. Em muitas situações não é viável a aplicação do NEO-PI-R, e para isto foram desenvolvidas algumas versões simplificadas deste questionário. Dentre estes podem ser destacados: *Big-Five Inventory* (BFI-44) com 44 questões e tempo estimado de 5 minutos para ser completado (John e Srivastava, 1999), apresentado no Apêndice A; o *NEO Five-Factor Inventory* (NEO-FFI) (Costa e McCrae, 1989) com 60 questões e tempo estimado de 10 minutos e o *Trait Descriptive Adjectives* (TDA) (Goldberg, 1990) com 100 questões e tempo previsto de 15 minutos. Duas versões mais simplificadas ainda foram desenvolvidas por Gosling et al. (2003), contendo 5 e 10 questões, respectivamente.

É natural que quanto maior o número de questões do questionário aplicado, maior será a precisão do levantamento de perfil realizado, sendo que os pesquisadores têm procurado dosar a dificuldade de aplicação de questionários, em função dos objetivos desejados, com a pesquisa e o volume de usuários disponíveis para a realização de seus experimentos. Duas estratégias são utilizadas para a identificação da personalidade: autoavaliação e observação. No caso da autoavaliação, o indivíduo preenche o questionário sobre si mesmo, é um processo de fácil interpretação e de fácil aplicação. No caso da observação, um indivíduo, preenche o questionário interpretando a personalidade de um outro indivíduo de sua proximidade e conhecimento, sendo que muitas vezes são utilizados vários observadores para compor uma medida média dos traços de personalidade do indivíduo observado. Esta situação de observação é conhecida como Percepção da Personalidade (Peabody, 1970).

2.2 Computação da Personalidade

O elo de ligação entre a Personalidade e a Computação pode ser verificado pela relação entre os traços de personalidade e a utilização da tecnologia. O conceito de Computação da Personalidade (tradução livre para *Personality Computing*) é apresentado por Vinciarelli e Mohammadi (2014) como a disciplina que aborda três problemas fundamentais, originários dos diferentes aspectos do modelo de lentes (Brunswik, 1956), conforme apresentado na Figura 2.2:

- Reconhecimento Automático de Personalidade (APR)
- Percepção Automática da Personalidade (APP)
- Síntese Automática da Personalidade (APS)

Inicialmente desenvolvido para explicar como os seres vivos capturam informações do meio ambiente, o modelo de lentes de Brunswik foi utilizado para as características socialmente relevantes de externalização e atribuição entre seres humanos e entre humanos e máquinas. As

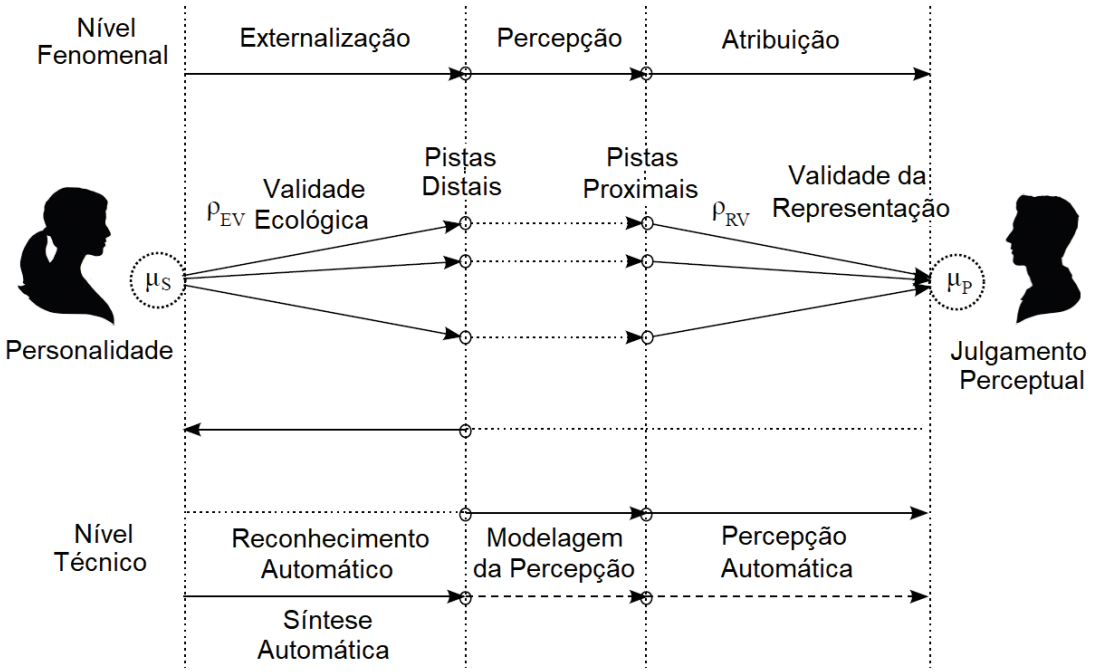


Figura 2.2: Personalidade e Modelo de Lentes de Brunswik

Fonte: Adaptado de Vinciarelli e Mohammadi (2014)

seções seguintes descrevem este modelo e sua relação com os problemas fundamentais tratados pela Computação da Personalidade.

2.2.1 Reconhecimento Automático de Personalidade (APR)

O Reconhecimento Automático de Personalidade (APR - *Automatic Personality Recognition*) está voltado ao processo de externalização e relacionado a tarefa de inferir personalidades, autoavaliadas previamente, a partir de pistas distais (distantes) detetáveis automaticamente. Esta tarefa é denominada Reconhecimento pois tem como objetivo realizar a inferência dos traços de personalidade autoavaliados previamente (utilizando os questionários abordados na Seção 2.1.2, por exemplo), considerados como traços verdadeiros da personalidade do indivíduo (Rammstedt e John, 2007). Em muitos casos a APR utiliza abordagens da Computação Afetiva, Processamento de Sinal Social e outras áreas ligadas a identificação de fenômenos emocionais e sociais através de evidências de comportamento detectáveis automaticamente.

Qualquer medida de covariância entre os traços de personalidade e as pistas distais, utilizando o coeficiente de *Spearman* (ρ) (Myers e Sirois, 2006) por exemplo, é conhecido como "Validade Ecológica" das pistas. Em pesquisa computacional os estudos de covariância são frequentemente atribuídos a seleção de características para reconhecimento de padrões. As abordagens de APR apresentadas na literatura incluem uma grande variedade de pistas distais envolvendo:

Reconhecimento em Texto - A psicologia da linguagem mostra que a escolha das palavras que falamos ou escrevemos não é direcionada somente pelo significado que queremos demonstrar, mas também por fenômenos psicológicos que queremos expressar tais como relações, atitudes, emoções e traços de personalidade (Tausczik e Pennebaker, 2010);

Comunicação Não Verbal - A Psicologia sugere que a comunicação não verbal é uma externalização da personalidade (Ekman et al., 1980), podendo ser utilizada para reconhecimento automático da personalidade, utilizando as pistas identificadas pelo comportamento não verbal, por exemplo. Este reconhecimento pode ser realizado utilizando os aspectos não verbais da fala, distâncias interpessoais, e combinação multimodal e movimentos corporais;

Mídias Sociais - As mídias sociais fornecem uma grande fonte de informações para a identificação da personalidade, visto ser um importante canal onde as pessoas interagem entre si, de forma livre. Além das informações textuais, podem ser obtidas referências sobre idade, localização geográfica, etnia, religião, educação, trabalho, amigos, dentre outras;

Dispositivos Móveis e Vestíveis - Os dispositivos móveis fazem parte do cotidiano das pessoas, com destaque para os *smartphones*. Além das funções básicas de ligações de voz, vídeo e mensagens, estes dispositivos agregam um conjunto de sensores, como os de localização, proximidade e aceleração, que podem ser usados como fontes de medidas da vida de seus usuários (Raento et al., 2009). Estas informações poderiam se utilizadas, ao menos em parte, para identificação de perfil de personalidade;

Jogos de Computador - Os jogos de computador representam uma importante indústria, e iniciativas na literatura estão propondo abordagens para inferir os traços de personalidade a partir de estratégias e opções adotadas pelos jogadores.

2.2.2 Percepção Automática de Personalidade (APP)

O Modelo de Lentes realiza distinção entre as pistas distais e proximais, visto que o processo de percepção é a representação mental de algo que é percebido (Thorndike, 2013). Nós não temos a percepção da energia das ondas acústicas quando escutamos uma música em um alto falante, mas sentimos a intensidade do volume de som que chega em nossos ouvidos. Neste caso a energia elétrica no alto falante é a pista distal e o ruído sonoro é a pista proximal. A pista proximal ativa o processo de atribuição, ou seja, o desenvolvimento do processo de julgamento para atribuição dos traços de personalidade que um observador atribui a um outro indivíduo que esteja observando.

A "Percepção Automática da Personalidade" é a tarefa de realizar a inferência dos atributos de personalidade, observados em um indivíduo, a partir das pistas proximais. Ao contrário da APR, o objetivo da APP não é a identificação da "verdadeira" personalidade do indivíduo, mas sim, a personalidade que é atribuída a este indivíduo por outros. Na APR a referência para o perfil de personalidade é obtida através de uma autoavaliação, ou seja, por informações que um indivíduo fornece sobre si mesmo, com a aplicação de um formulário de identificação de personalidade. As abordagens de APP em geral utilizam a predição da personalidade utilizando a média das trilhas atribuídas por um conjunto de avaliadores.

Como exemplos de utilização da APP podem ser destacados:

Paralinguagem - A paralinguagem inclui os aspectos associados à fala, tais como o tom, o ritmo e o volume, que permitem a identificação de outras características na fala, além das palavras pronunciadas. Conforme verificado por Ekman et al. (1980), os elementos da análise da fala, incluindo os elementos da paralinguagem, tem uma forte correlação em como as pessoas realizam os seus julgamentos;

Comportamento Não Verbal - O comportamento não verbal está associado aos elementos que fornecem indicativos sobre a personalidade tais como gestos e expressões faciais, que podem influenciar no julgamento da personalidade alheia;

Mídias Sociais - Em contraste aos estudos de APR em mídias sociais, as pesquisas em APP nesta área, são bem reduzidas. Nesta área os traços de personalidade são auferidos pela análise, por parte dos julgadores, das informações depositadas pelos usuários, como por exemplo as imagens indicadas como favoritas, ou as imagens de perfil.

2.2.3 Sintetização Automática de Personalidade (APS)

Uma das maiores descobertas do estudo da cognição social é de que os indivíduos naturalmente associam características socialmente relevantes, tais como os traços de personalidade, aos indivíduos com os quais se relacionam (Uleman et al., 2008). Este fenômeno não se aplica somente às pessoas, mas também a qualquer dispositivo que exiba características humanas (Reeves e Nass, 1998).

A APS pode ser descrita como a tarefa de gerar automaticamente as pistas distais necessárias para provocar a atribuição dos traços de personalidade desejados. Utilizando o modelo de lentes de Brunswik (Figura 2.2), a APS abrange os processos de externalização e atribuição. No caso da externalização, as pistas não são geradas por humanos, mas sim por qualquer dispositivo que possa apresentar comportamentos similares aos humanos. O processo de atribuição envolve observadores humanos que associam, mesmo inconscientemente, os traços de personalidade gerados pelo dispositivo. O objetivo principal do processo é que os traços associados pelos observadores humanos correspondam aos desejados pelos projetistas dos dispositivos.

Como aplicações da APS temos:

Baseado em Fala - O processo de APS baseado em fala consiste em realizar a sintetização da fala humana, com imposição de traços de personalidade determinados. Conforme verificado por Nass e Lee (2001), as vozes sintetizadas podem manifestar a personalidade;

Agentes Inteligentes - Um agente inteligente, é tipicamente uma solução de software, que adota a melhor ação possível mediante uma determinada situação. Estes mecanismos estão presentes nos dispositivos de busca na internet, sistemas de resposta audível e "avatares" em IHC, por exemplo. A APS pode ser utilizada para gerar traços artificiais de personalidade nestes agentes;

Robôs - Tal qual no caso dos Agentes Inteligentes, a utilização da APS em robótica permite gerar artificialmente traços de personalidade nos robôs por meio de expressões faciais, corporais e paralinguagem na fala.

2.2.4 Considerações e Restrições

A Computação da Personalidade apresenta os diferentes aspectos de identificação e sintetização da personalidade utilizando processos automatizados. Como restrição de escopo da presente pesquisa, somente serão tratados na sequência, os aspectos relacionados ao Reconhecimento Automático da Personalidade (APR) a partir de informações de texto, e as tecnologias envolvidas neste processo.

Neste universo, as técnicas de identificação que utilizam estratégias de classificação e regressão, em geral binárias, buscam a descoberta das pistas distais para inferência dos traços

de personalidade. Outro aspecto a ser considerado nestas estratégias observadas nos trabalhos publicados q é de que as dimensões de personalidade, OCEAN no caso do BIG FIVE, não são correlacionadas e são independentes (Saucier e Srivastava, 2015). Apesar de estudos mais recentes procurarem demonstrar os indícios de correlação entra estas dimensões, por restrição metodológica, neste presente trabalho de pesquisa a identificação do perfil de personalidade adotará a premissa simplificatória de que as dimensões não são correlacionadas.

Os ensaios relacionados a soluções de APR tem utilizado bases de referência previamente mapeadas para levantamento de métricas sobre a validade destas soluções. Os resultados apresentados estão também baseados na fidedignidade destes mapeamentos, que são amparados pelos histórico dos autores que os realizaram, no caso das bases de dados ESSAYS (Pennebaker e King, 1999) e *myPersonality* (Kosinski et al., 2015). Esta premissa metodológica também foi assumida neste projeto de pesquisa, que utiliza nos seus ensaios estas duas bases de dados.

2.3 Educação e Personalidade

Conforme verificado por Crozier (2013), a maioria dos estudos que tratam da relação entre a educação e a personalidade, relata influências da personalidade dos alunos no processo de aprendizagem, mas não há paradigma para vincular esses estudos. A personalidade é frequentemente considerada um obstáculo nos programas educacionais. Enquanto os professores preferem tratar todos os alunos da mesma forma, as diferenças entre estes são consideradas problemas no ensino e educação, e não como recursos a serem desenvolvidos De Raad e Schouwenburg (1996). Considerar as diferenças é encontrar situações de aprendizagem ótimas para cada aluno, buscando uma educação individualizada (Perrenoud, 2001).

O sucesso do processo de aprendizagem tradicionalmente foi associado ao conceito de inteligência do aluno, sendo esta descrita por Antunes (1998) como a capacidade cerebral pela qual conseguimos penetrar na compreensão das coisas escolhendo o melhor caminho. Em oposição a ideia de que a inteligência é algo monolítico e que cada indivíduo a possui em diferentes níveis, Gardner e Hatch (1989) com sua "Teoria das Inteligências Múltiplas", propõe que a vida humana requer o desenvolvimento de vários tipos de inteligência. Esta teoria enfatiza a necessidade de uma educação diferenciada que esteja preparada para atender a estas diferenças individuais, procurando desenvolver o potencial de cada indivíduo, utilizando metodologias e estratégias adequadas para a estimulação das diferentes inteligências, que de acordo com Gardner, podem ser classificadas como:

- intrapessoal ;
- linguística;
- corporal-cinestésica;
- interpessoal;
- lógico-matemática;
- musical;
- naturalista e
- espacial.

Existem boas evidências que demonstram que a personalidade ocupa um papel importante no aprendizado e a sua influência pode ser demonstrada tanto em ambientes de laboratório como na sala de aula em si (Eysenck, 1978). Os traços básicos de personalidade podem indicar uma estratégia metodológica e a maneira pela qual o indivíduo processa as informações, e podem ser utilizados como uma ferramenta de medida para o ensino (Messick, 1984). Segundo De Raad e Schouwenburg (1996), os fatores *Conscientiousness*, *Extraversion* e *Openness* são educacionalmente relevantes. Busato et al. (1998) verificaram que: a *Extraversion* tem correlação

positiva com significados, aplicação e reprodução dirigidos; *Conscientiousness* também tem associação positiva com estes mesmos componentes, mas correlaciona-se negativamente com atividades não dirigidas; *Openness* tem correlação direta com significado dirigido e negativamente com o não dirigido; o *Neuroticism* tem correlação positiva com estilo de aprendizagem dirigido e negativamente com significado dirigido; *Agreeableness* é associado positivamente com reprodução dirigida.

2.3.1 Desempenho Acadêmico

O estudo do Desempenho Acadêmico tem como objetivo entender como o estudante, o professor e a instituição procuram alcançar os objetivos educacionais. Além da preocupação inerente das instituições em procurar com que um número cada vez maior de estudantes alcancem estes objetivos dentro dos prazos desejados, também existe o reflexo destes indicadores nos resultados apresentados pela instituição para a comunidade, como índices de aprovação, evasão e coeficientes de rendimento global. Os resultados apresentados por uma instituição podem ter reflexos nos valores recebidos dos seus patrocinadores, tanto nas instituições públicas quanto nas instituições privadas. No caso do Brasil, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) realiza a avaliação do ensino superior para a instituição como um todo, para os cursos e para os estudantes. A avaliação específica do desempenho dos estudantes, nas instituições de ensino superior públicas e privadas é realizada por meio do Exame Nacional de Desempenho de Estudantes (ENAD).

De acordo com Magalhães et al. (2009), o Desempenho Acadêmico é influenciado por: características de personalidade; características sócio-demográficas; fatores contextuais, ligados à instituição de ensino; concepções de aprendizagem dos alunos; processo de adaptação à universidade por parte dos estudantes e as experiências no contexto acadêmico, como o *stress*.

Os testes de habilidade cognitiva foram desenvolvidos especificamente para avaliar a habilidade nata do indivíduo ao passo que os testes de aproveitamento tem sido desenvolvidos para prever as diferenças individuais no aprendizado (Brody, 2000). Estudos contemporâneos tem enfatizado a importância de observar os indicadores de Desempenho Acadêmico não relacionados com cognição e nem com a habilidade (Chamorro-Premuzic e Furhnam, 2006). Enquanto os testes de habilidade indicam o que uma pessoa pode fazer, os fatores não cognitivos podem fornecer importantes informações sobre o que uma pessoa irá fazer. Conforme observado por Furnham et al. (2009), as diferenças individuais, em comportamento normal, podem ser classificadas em termos de dimensões diferentes e ortogonais, como no Modelo BIG FIVE, descrito na Seção 2.1.1, refletindo as diferenças individuais de forma estável e as preferências que determinam os padrões característicos de cada indivíduo em termos de pensamento, emoção e comportamento, representando medidas agregadas do comportamento individual e que podem ser avaliados por questionários, conforme descrito na Seção 2.1.2.

A literatura apresenta estudos que verificam a influência da personalidade no Desempenho Acadêmico. A dimensão *Conscientiousness* está diretamente relacionada com os diferentes resultados acadêmicos: provas; redações; avaliação continuada e dissertações supervisionadas (Heaven et al., 2007; O'Connor e Paunonen, 2007). O *Neuroticism* representa um indicador negativo em muitos resultados, particularmente quando os estudantes são avaliados por exame final ou outros métodos estressantes (Laidra et al., 2007). A dimensão *Openness* esta significativamente associada com os resultados em alguns estudos, mas não em outros (Chamorro-Premuzic e Furnham, 2003). Esta ambiguidade também ocorre com a dimensão *Extraversion* (Wolf e Ackerman, 2005), que aparenta ser mais relacionada positivamente durante o ensino básico, mas negativamente no ensino superior. Segundo Poropat (2009) as dimensões *Conscientiousness*

e *Openness* apresentam associações mais consistentes com o Desempenho Acadêmico. De qualquer modo, a relação entre os traços de personalidade e o Desempenho Acadêmico são evidentes, indicando a importância da utilização desta identificação pelos educadores.

2.3.2 Educação a Distância

A educação formal depende de vários fatores, como o estilo utilizado pelo instrutor e a estratégia de entrega das atividades, que podem influenciar os resultados alcançados no processo (Vonderwell e Zachariah, 2005). Cada evento educacional pode ser considerado um evento único e para maximizar estes resultados, os educadores devem procurar reconhecer e se adaptar às particulares destes eventos. É esperado que em ambientes onde o comportamento dos indivíduos é menos restrito, os traços de personalidade possam refletir melhor os fenômenos de comportamento (Mischel, 2013). Isto pode ser notado em ambientes de Educação a Distância (EaD), onde os estudantes tem mais liberdade para atender às suas expectativas sociais, tendo mais liberdade sobre a natureza e frequência das interações sociais (Varela et al., 2012).

2.3.3 Estilo de Aprendizagem

Muitas pessoas consideram o processo de aprendizagem como algo natural e independente de assistência, sendo concluído na fase adulta da vida. De acordo com Skinner (1982) a aprendizagem é basicamente uma mudança de comportamento. Esta ocorre quando alguém demonstra saber algo que não sabia antes. O processo de aprendizagem pode ser descrito como a maneira pela qual as pessoas adquirem, armazenam e utilizam conhecimento.

O estilo de aprendizagem é um conjunto de características pessoais, desenvolvidas biologicamente e durante o crescimento, que faz com que o mesmo método de ensino seja eficiente para alguns e ineficiente para outros (Dunn et al., 2002). De acordo com Cerqueira (2000), o estilo de aprendizagem é muito importante para os professores, porque influencia em sua maneira de ensinar, uma vez que os professores tendem a ensinar da maneira que gostariam de aprender, ou seja, seguindo seu estilo de aprendizagem e não o estilo dos alunos.

Essencialmente, o estilo de aprendizagem possui três componentes: i) maneira com que se processa a informação; ii) seleção dinâmica de estratégias de aprendizagem; e iii) própria percepção da pessoa com respeito a sua aprendizagem.

Conforme Dunn et al. (1977), a orientação da aprendizagem de uma pessoa é, talvez, o determinante mais importante de sua realização educacional. Quanto maior sua aderência com o método pedagógico utilizado, maior a chance de sucesso (Delahaye e Thompson, 1991). Como consequência, existem instrumentos que procuram medir os estilos de aprendizagem. Muitos autores pesquisaram o conceito de estilos de aprendizagem resultando em dezenas de modelos (Hayes e Allinson, 1988; Coffield et al., 2004). A Figura 2.3 ilustra estes modelos agrupados por famílias. Os estilos de aprendizagem indicados em negrito representam os 13 estilos principais, sendo os demais considerados derivados destes.

Enquanto uma série de estudos foram realizados na área do estilo de aprendizagem e como este pode colaborar na utilização de estratégias individualizadas para os alunos, no processo de ensino, Kirschner (2017) propõe que o “mito” dos estilos de aprendizagem deixe de ser propagado. O autor indica que ainda não existem evidências concretas de que o agrupamento dos alunos de acordo com o seu estilo encontra suporte adequado nos objetivos educacionais. Apesar disto, novas pesquisas estão sendo conduzidas, como Silva (2017) que realizou um estudo empírico em que o modelo de Felder-Silverman foi utilizado na adaptação das interfaces para os estudantes, comprovando a aplicabilidade deste modelo.

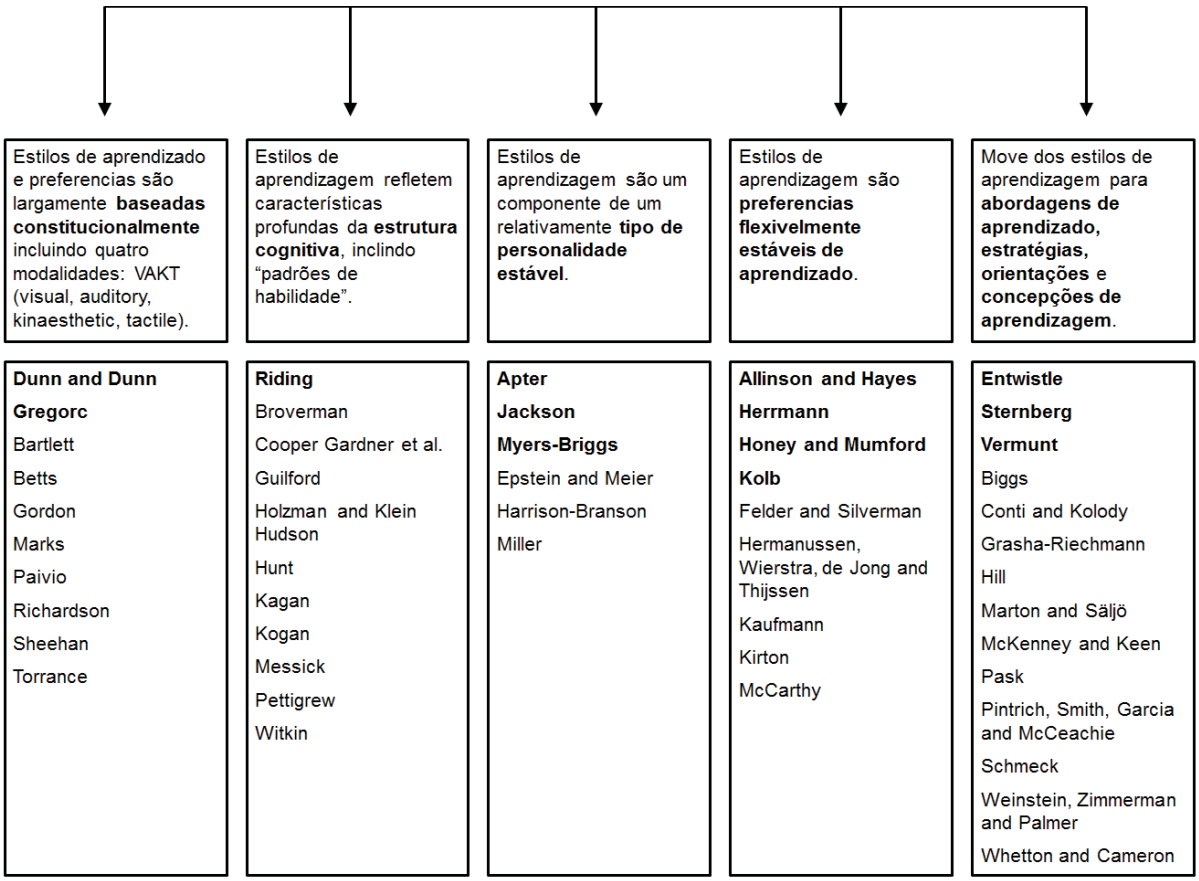


Figura 2.3: Lista de Estilos de Aprendizagem

Fonte: Coffield et al. (2004)

A estreita relação entre a personalidade e os estilos de aprendizagem tem sido verificada por diversos estudos (Miller, 1991; Busato et al., 1998; Cassidy, 2004; Komarraju e Karau, 2005). A partir da identificação dos traços de personalidade, poderiam ser realizadas inferências sobre o estilo de aprendizagem do aluno, propiciando a aplicação de estratégias de ensino individualizadas que sejam mais adequadas a cada um destes.

2.4 Reconhecimento Automático a partir de Texto

A utilização dos questionários de avaliação da personalidade na maioria das vezes são extensos e intrusivos (Gosling et al., 2003), o que dificulta a extração intencional destas informações. Outras formas mais amigáveis, menos árduas e menos intrusivas, de extração de personalidade, tem sito investigadas com o objetivo de reduzir o impacto causado pela utilização dos questionários, sendo uma delas, a investigação dos traços de personalidade identificados no texto.

A maior parte dos estudos de identificação das características psicológicas do indivíduo, a partir das informações textuais, tem sido direcionada para a detecção de polarização (positiva, negativa ou neutra) (Pang et al., 2008), processo este conhecido como "Identificação de Emoções". O processo de identificação da personalidade, e consequente identificação das subjetividades registradas em um texto, de forma computacional, é uma tarefa complexa e exaustiva. Conforme (Munezero et al., 2014), a subjetividade do ser humano tem forte relação com o afeto, com os

sentimentos, com as emoções e com as opiniões. Um modelo computacional utilizado para o reconhecimento de personalidade (APR), denominado na presente pesquisa como BRC, é composto por três componentes distintos, conforme o diagrama apresentado na Figura 2.4:

Base de Dados - Base de avaliação do modelo, previamente rotulada;

Representação - Método de representação do texto, por meio de características;

Classificação - Técnica de classificação utilizada.

$$BRC = [\text{Base de Dados, Representação, Classificação}] \tag{2.2}$$

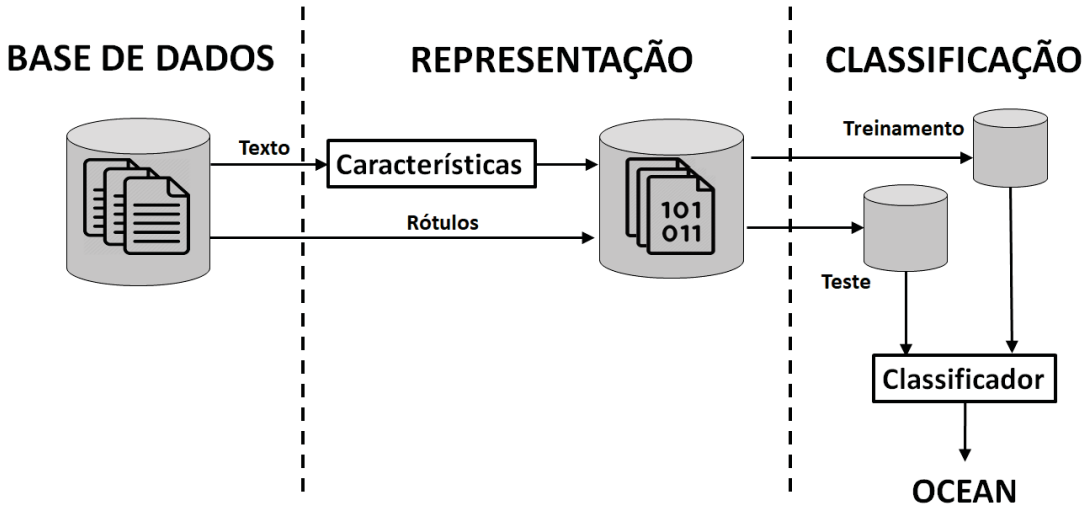


Figura 2.4: Modelo BRC

Este modelo está baseado nos conceitos de Aprendizagem de Máquina, em que os padrões identificados no texto, são comparados com os padrões identificados em uma grande base de dados, chamada base de treinamento. A base de treinamento é composta por diversas amostras previamente classificadas manualmente, relacionando o perfil de personalidade e conjuntos de textos redigidos por voluntários. A partir dos padrões de semelhança encontrados e utilizando as diversas técnicas de classificação, é então realizada a inferência do perfil de personalidade do autor de um texto que foi submetido a este processo de classificação. As seções 2.5, 2.6 e 2.7 descrevem as funcionalidades destes componentes, os modelos mais utilizados e as opções e limitações definidas para estes componentes, na presente pesquisa.

2.5 Base de Dados

As bases de dados utilizadas nos experimentos de identificação de personalidade a partir do texto são formadas por um conjunto de textos livres, elaborados por voluntários, que são mapeados de acordo com o perfil de personalidade destes voluntários. Os voluntários são submetidos a algum processo de identificação de personalidade conforme um dos modelos de personalidade que no caso das bases aqui abordadas é o BIG FIVE. Desta forma, é obtida a relação entre o voluntário, o texto por este produzido e os valores OCEAN que correspondem a sua personalidade, identificada previamente de forma manual. Este conjunto de informações permite

a realização de experimentos para avaliação dos modelos de identificação de personalidade propostos pelos pesquisadores, visto possuírem os textos para serem explorados e os rótulos referentes ao perfil de personalidade dos autores dos textos. As bases podem ser divididas e utilizadas nos processos de classificação como bases de treinamento, bases de validação e bases de teste. Nesta seção são apresentados as duas bases mais utilizados nos experimentos investigados, relacionados ao modelo BIG FIVE, que são as bases ESSAYS e *myPersonality*, sendo que a estrutura e composição destas bases são detalhadas na Seção 6.4.

2.5.1 Base ESSAYS

A base denominada ESSAYS é composta por um conjunto de textos produzidos por 2467 indivíduos, coletados por James W. Pennebaker (Pennebaker e King, 1999) durante o desenvolvimento do léxico LIWC (seção 2.6.3). Estes textos foram coletados entre os anos de 1997 e 2004, correspondendo a textos produzidos livremente por alunos de psicologia, pacientes em tratamento psicológico, alunos de turma de verão e psicólogos sociais renomados, sendo que todos os textos estão no idioma inglês. Foi solicitado aos voluntários a realização de uma redação durante cerca de 20 minutos sobre um tópico específico, sendo este procedimento repetido durante 10 dias consecutivos.

A obtenção da classificação manual do perfil de personalidade dos voluntários foi realizada com a aplicação do formulário BFI-44 (John e Srivastava, 1999), descrito na Seção 2.1.2. O resultado da aplicação destes formulários foram avaliados pelos pesquisadores, ocorrendo posteriormente o registro de um rótulo "*yes*" ou "*no*", na base de dados, correspondente a cada uma das dimensões OCEAN dos diversos voluntários. O formato de apresentação das informações desta base está apresentado na Seção 6.4.1.

2.5.2 Base *myPersonality*

O Projeto *myPersonality* oferece um *site* na *internet* que permite a identificação do perfil de personalidade de voluntários por meio da aplicação de testes psicométricos *on-line* e pela busca dos textos inseridos por estes mesmos usuários em seus correspondentes perfis na rede social *Facebook* (Kosinski et al., 2015). Composto por uma gama heterogênea de participantes, os dados correspondentes a esta identificação estão disponibilizados de maneira pública, na forma de uma tabela, denominada "*10,000 Facebook status updates of 250 users + personality + Facebook social network properties*" (Goldberg et al., 2006), a qual será referenciada nesta pesquisa como *myPersonality*.

Os voluntários que contribuíram com esta base são oriundos de grupos de diversas faixas de idade, culturas e perfis. Além das dimensões OCEAN, apresentadas na faixa de 1,00 a 5,00, bem como "yes" ou "no", são apresentadas informações adicionais de demografia, comportamento social, interesses, preferências, opiniões, obtidas dos perfis do *Facebook* dos voluntários que autorizaram a coleta destas informações. Apesar da base *myPersonality* oferecer estas características relacionadas a identificadores da rede social dos voluntários, por "restrição metodológica da presente pesquisa", somente foram utilizados nos experimentos realizados, as informações dos textos publicados pelos voluntários e a classificação prévia das dimensões OCEAN.

2.6 Representação

Um texto é por natureza uma informação primariamente não estruturada, ou seja, não é viável a realização de um processo de classificação de forma direta. Para isto o texto precisa ser submetido a um processo de representação por meio de características, para viabilizar a execução de modelos de classificação que permitam a identificação dos traços de personalidade do autor do texto. Com o processo de obtenção de características que representem o texto, a quantidade de recursos necessários para descrever uma grande quantidade de dados, como um texto, é reduzida. Além disto as informações de natureza textual são expressas como um vetor de valores numéricos que representem aquele texto em questão. Os vetores de natureza numérica são mais adequados para a aplicação de métodos de classificação. Este vetor numérico é obtido utilizando diferentes técnicas que visam representar o texto, de acordo com a ótica do modelo adotado.

Os principais métodos para obtenção de características estão associados a informações estatísticas sobre o texto. De uma forma simplista, pode ser realizado o levantamento da frequência de ocorrência de cada palavra no texto, e desta forma ser gerado um vetor que contenha esta informação. E este vetor é ampliado em função de novos textos que forem classificados na sequência. A utilização de técnicas de Processamento de Linguagem Natural (PLN) (Seção 2.6.5) pode reduzir o tamanho do vetor obtido. Outro método utilizado para a obtenção de características do texto é por meio da utilização de léxicos, que associam as palavras a certas categorias pré-definidas. Os léxicos permitem a classificação das palavras de acordo com algumas categorias, que podem incluir as emoções, no caso dos léxicos afetivos (Seção 2.6.2).

2.6.1 Fontes de Texto

A Psicolinguística, uma ciência que surgiu a partir da Linguística e da Psicologia, tem procurado entender e explicar a estrutura mental e os processos envolvidos no uso de uma língua. Um texto é uma representação formal deste uso da língua. As atividades de formulação textual compreendem: a escolha de informações que aparecerão no texto; a organização que se dá a estas informações e que é afetada por fatores diversos, questões cognitivas e argumentativas e a escolha e construção da forma linguística que servirá de veículo para tais informações e as estratégias argumentativas que as acompanham. Nesta formulação entram os elementos de todos os planos lexical, frasal e textual, bem como os níveis fonológico, morfológico, sintático, semântico, pragmático da língua (Travaglia, 2016). A presença de pistas que conduzam a inferência dos traços de personalidade no texto foi indicada pelas Hipóteses Léxicas (Allport, 1937) em que as mais relevantes diferenças individuais estão codificadas na linguagem, ou seja, presentes no texto.

Uma das fontes de texto utilizadas nos processos de identificação de personalidade é a utilização de textos formais, como redações e relatórios, onde o autor tem uma preocupação mais rígida com a forma e a gramática. Foi a partir deste tipo de texto que Pennebaker e King (1999) compilaram a base ESSAYS.

Outra fonte que vem sido muito utilizada nos estudos relacionados a identificação de personalidade são os textos obtidos de *Twitter*, *Facebook*, *Blogs* e outros meios similares. Um dos motivos da utilização deste tipo de fonte é a disponibilidade de grandes quantidades de informações textuais de forma *online*. Nestas fontes não ocorre uma grande preocupação com formalismo, sendo comum a presença de neologismos.

Nas atividades educacionais, com maior ênfase em EaD, podem ser obtidos textos, redigidos pelos alunos, a partir de atividades diversas solicitadas pelos professores. Neste tipo de

texto, além da preocupação com a forma, ocorre a presença de textos relacionados ao tema a ser discutido na tarefa, que muitas vezes não demonstram livremente os traços de personalidade.

2.6.2 Léxicos Afetivos

Os léxicos podem ser descritos como um conjunto de palavras disponíveis em um determinado idioma que as pessoas utilizam para realizar a expressão oral e escrita de suas ideias. O termo vocabulário está associado ao subconjunto de palavras do léxico, que um determinado indivíduo utiliza em suas verbalizações e expressões. Ou seja, o léxico está associado a um idioma e o vocabulário, a um indivíduo específico. Em termos computacionais, os léxicos são utilizados em sistemas computacionais na área de Processamento de Linguagem Natural (PLN) e devem possuir informações adequadas e codificadas para que o algoritmo ou programa desenvolvido possa compreendê-lo e executá-lo. O termo *corpus* é utilizado para associar um léxico que esteja disponível em forma eletrônica. Além de um simples conjunto de palavras, os léxicos podem trazer informações sobre sinônimos, antônimos, categorias sintáticas e outras informações úteis para o PLN.

Um Léxico Afetivo é um conjunto no qual as palavras estão associadas a emoções e condições afetivas, como o afeto, ânimo e sentimento (Ortony et al., 1987). As emoções são causadas por condições afetivas, considerando que nem todas as condições afetivas são emoções. São utilizadas principalmente para a investigação das emoções expressadas de forma escrita, ou oral, por um indivíduo.

2.6.2.1 ANEW

O léxico afetivo *Affective Norms for English Words*(ANEW) (Bradley e Lang, 1999) é um conjunto de palavras afetivas que tenham características de emoção, sendo dividido em três emoções: agradável a desagradável; calmo a exaltado e dominado a controlado. A avaliação das escalas foi realizada com a utilização de desenhos de bonecos, na escala denominada como *Self-Assessment Manikin* (SAM (Figura 2.5). Foram avaliadas e analisadas 1.044 palavras neste léxico.

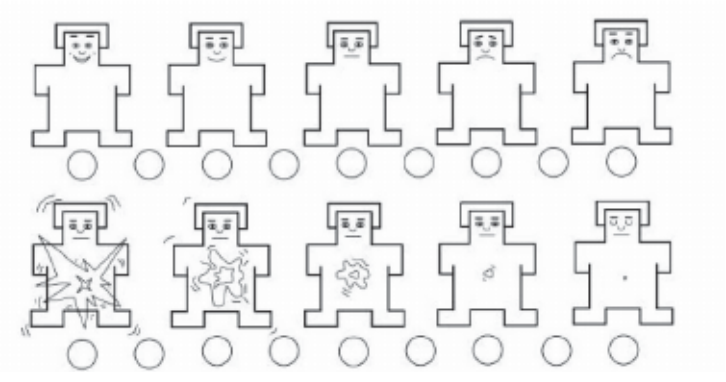


Figura 2.5: Exemplo de *Self-Assessment Manikin* (SAM)

Fonte: Nijboer et al. (2009)

Este léxico foi adaptado para a língua portuguesa por Kristensen et al. (2011), com a denominação de *ANEW-Br*. Além da tradução literal, foi realizada a tradução reversa para a língua inglesa, bem como verificação posterior por um grupo de avaliadores independentes para validação do processo.

2.6.2.2 SentiWordNet

O léxico afetivo *SentiWordNet* foi desenvolvido por Esuli e Sebastiani (2007) (versão 1.0) e disponibilizado para acesso público. É um dicionário desenvolvido para apoiar iniciativas de classificação de textos baseado em sentimentos e opinião (Pang et al., 2008). Este recurso é o resultado da categorização das palavras do léxico *WordNet* (Miller, 1995) de acordo com valores "positivo", "negativo" e "neutro". O *WordNet*, por sua vez, é um dicionário para a língua inglesa contendo mais de 166.000 verbetes, que foi compilado com o objetivo de ser utilizado por processos computacionais. É um recurso disponibilizado *online* pela *Princeton University* (EUA), que agrega substantivos, adjetivos, advérbios e adjetivos organizados em um conjunto de sinônimos e relações semânticas.

2.6.2.3 Wordnet-Affect

O *WordNet-Affect*, é uma extensão afetiva do léxico *wordNet Domains* (Magnini e Cavaglia, 2000), que por sua vez foi desenvolvido a partir do *WordNet* (Miller, 1995). Foi concebido para ser um conjunto de conceitos afetivos correlacionados a palavras afetivas, contendo cerca de 4.700 palavras. A variante *Wordnet-AffectBR* foi adaptada por Pasqualotti e Vieira (2008) para a língua portuguesa.

2.6.3 LIWC

A ferramenta LIWC (*Linguistic Inquiry and Word Count*) é uma ferramenta de *software* que permite o cálculo da frequência de ocorrência de diferentes categorias de palavas em um conjunto de textos fornecidos (Pennebaker et al., 2001). Esta ferramenta tem sua funcionalidade baseada no dicionário LIWC, que é um léxico contendo cerca de 4.500 palavras previamente mapeadas em uma ou mais categorias diferentes, dentre as dezenas de categorias existentes neste dicionário. Na sequência deste trabalho, o termo LIWC quando utilizado isoladamente, será uma referência ao léxico semântico em si. Durante duas décadas, Dr. James Pennebaker realizou pesquisas sobre a relação entre linguagem, psicologia e saúde. Em conjunto com Martha Francis e Roger Booth, desenvolveu a ferramenta LIWC e o correspondente dicionário. A primeira versão do LIWC surgiu em 1993 como na forma de um estudo exploratório da linguagem e como forma de divulgação do trabalho (Pennebaker et al., 1993).

A segunda versão, denominada LIWC2001 (Pennebaker et al., 2001), bem como a terceira, que foi chamada LIWC2007 (Pennebaker et al., 2007) foram atualizações do trabalho original e expansões no dicionário, além de melhorias no projeto da ferramenta de *software*. Na mais recente versão disponível, o léxico LIWC2015 (Pennebaker et al., 2015) recebeu uma alteração significativa em seu dicionário, bem como na ferramenta em si, permitindo a utilização de diversos dicionários além do LIWC2015, tal como o LIWC2001 ou LIWC2007. Além disto, podem ser utilizados com dicionários de outras linguagens, como espanhol, alemão, italiano e português (Balage Filho et al., 2013), dentre outras.

O desenvolvimento do léxico LIWC foi fruto de muitos anos de trabalho e originalmente surgiu da ideia de identificar grupos de palavras associadas a dimensões cognitivas e emocionais básicas, frequentemente estudadas na psicologia da personalidade. Neste processo, foram realizados experimentos estatísticos válidos, mostrando a correlação entre o estilo linguístico e a personalidade (Pennebaker e King, 1999; Chung e Pennebaker, 2008; Tausczik e Pennebaker, 2010). Na versão atual (LIWC2015), o léxico agrega números, pontuação e até frases curtas, permitindo aproximar a análise da linguagem utilizada em redes sociais, permitindo a análise de alguns *emoticons* textuais, como por exemplo ":)" e acrônimos ou gírias como "b4" (*before*).

A criação do dicionário é um processo que envolveu diversas etapas. Em uma primeira etapa, palavras obtidas de diversos dicionários tradicionais e afetivos foram agregadas, e a categorização inicial foi realizada por um grupo de juízes que individualmente fizeram uma seleção. Na sequência, em sessões de *brain-storming*, grupos de juízes fizeram uma segunda seleção e ajustes na classificação inicial. Em uma fase seguinte de avaliação, os juízes realizaram a avaliação conjunta de cada palavra para verificar se esta permaneceria ou não nas categorias a qual foi associada, e em caso de não haver consenso, a palavra foi eliminada. Em uma próxima fase, com a construção do dicionário inicial, partindo da categorização feita pelos juízes, texto de diversas fontes foram analisadas com este dicionário, utilizando a ferramenta MEH (*Meaning Extraction Helper*) (Boyd, 2014), verificando a frequência de ocorrência de cada palavra dentro das diversas bases analisadas. As palavras que não tiveram ocorrido ao menos uma vez em todas as bases, foram descartadas.

Na fase seguinte, com a utilização do MEH, palavras com alta frequência de ocorrência em outras bases foram agregadas, com mais uma rodada de julgamento. Uma etapa de avaliação psicométrica foi realizada, com a separação de cada categoria com sua palavras constituintes. Cada palavra foi quantificada como um percentual do total de palavras para os cerca de 180.000 arquivos textos de 5 *corpora*, com uma faixa de 231 milhões de palavras. Um outro grupo de juízes revisaram a lista e escolheram as palavras que deveriam permanecer. Uma etapa de refinamento, que consistiu na repetição de todas a fases anteriores, foi então realizada. As categorias em que as palavras foram categorizadas, tendo como referência o léxico LIWC2015, foram separadas nos seguinte grupos e categorias:

Métricas - Número médio de palavras, palavras com mais de seis letras, palavras no dicionário;

Funcionais - total de funcionais, total de pronomes, pronomes pessoais, primeira pessoa no singular, primeira pessoa no plural, segunda pessoa, terceira pessoa no singular, terceira pessoa no plural, pronome impessoal, artigo, preposição, verbo auxiliar, advérbio, conjunção, negação ;

Outras Gramáticas - verbos comuns, adjetivos, comparações, interrogações, números, quantificadores;

Afetivos - total de afetivos, emoção positiva, emoção negativa, ansiedade, raiva, tristeza;

Sociais - total de sociais, família, amigos, referências femininas, referencias masculinas;

Cognitivos - total de cognitivos, discernimento, causa, discrepância, tentativa, certeza, diferenciação;

Perceptivos - total de perceptivos, ver, escutar, sentir;

Biológicos - total de biológicos, corpo, saúde, sexualidade, ingestão;

Necessidades - total de necessidades, , afiliação, realização, potência, recompensa, risco;

Temporais - foco passado, foco presente, foco futuro;

Relatividade - movimento, espaço, tempo;

Preocupações Pessoais - trabalho, lazer, lar, dinheiro, religião, morte;

Linguagem Informal - total de informais palavrão, linguagem da internet, aprovação, não fluência,pausa;

Pontuação - pontuação total, ponto, vírgula, dois pontos, ponto e vírgula, interrogação, exclamação, barra, aspas, apóstrofos, parêntesis, outras pontuações.

Esta lista de categorias demonstra a grande faixa de características que podem ser extraídas de um texto utilizando este recurso. Tendo como exemplo o pequeno texto de entrada "*likes the sound of thunder.*", após o processamento pelo LIWC2015, são obtidas as categorias apresentadas na Tabela 2.4.

Tabela 2.4: Categorias LIWC

Categoria	Valor	Categoria	Valor
Número de palavras	5	Processo afetivo	20%
Palavras por sentença	5	Emoção positiva	20%
Palavras de seis letras ou mais	20%	Percepção	40%
Palavras no dicionário	100%	Escutar	40%
Total de funções	40%	Pontuação total	20%
Artigo	20%	Ponto	20%
Preposição	20%		

2.6.4 MRC

O dicionário denominado *MRC Psycholinguistic Database: Machine Usable Dictionary* (Wilson, 1988) contém cerca de 150.000 palavras em inglês mapeadas de acordo com 26 categorias linguísticas e psicolinguísticas. O dicionário ocupa cerca de 11MB. Além da base de dados, são disponibilizados três programas utilitários para acesso à base, denominados "DICT", "GETENTRY" e "PSYCHDICT". A Tabela 2.5 ilustra as propriedades descritas neste dicionário, bem como as frequências de ocorrências destas.

2.6.5 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é a área da Computação e da Inteligência Artificial que trata das interações entre humanos e computadores utilizando linguagem natural. Muitos modelos de identificação de personalidade utilizam técnicas de PLN para preparação do texto a ser analisado, sendo que algumas destas estão descritas nesta seção. Em função do modelo de identificação adotado, pode ser interessante, ou não, a aplicação de algumas, ou de diversas das técnicas descritas, visto que o PLN pode representar a eliminação de alguma informação que possa ser importante para o modelo em questão.

2.6.5.1 Normalização

A normalização é o processo de preparar um texto livre para ser processado. Um dos processos da normalização é a separação do texto em unidades menores. O primeiro passo é a separação do texto em frases (*sentence tokenize*), e depois a separação de frases em palavras (*word tokenize*).

Tendo como exemplo o texto:

$$Texto = \text{'Exemplo de texto simples. Segunda frase.'} \tag{2.3}$$

Tabela 2.5: Propriedades Descritas no Dicionário MRC

	Nome	Propriedade	Ocorrências
1	NLET	Number of letters in the word	150.837
2	NPHON	Number of phonemes in the word	38.438
3	NSYL	Number of syllables in the word	89.402
4	K-F-FREQ	Kucera and Francis written frequency	29.778
5	K-F-NCATS	Kucera and Francis number of categories	29.778
6	K-F-NSAMP	Kucera and Francis number of samples	29.778
7	T-L-FREQ	Thorndike-Lorge frequency	25.308
8	BROWN-FREQ	Brown verbal frequency	14.529
9	FAM	Familiarity	9.392
10	CONC	Concreteness	8.228
11	IMAG	Imagery	9.240
12	MEANC	Mean Colorado meaningfulness	5.450
13	MEANP	Mean Paivio meaningfulness	1.504
14	AOA	Age of acquisition	3.503
15	TQ2	Type	44.976
16	WTYPE	Part of speech	150.769
17	PDWTYPE	PD part of speech	38.390
18	ALPHSYL	Alphasyllable	15.938
19	STATUS	Status	89.550
20	VAR	Variant phoneme	1.445
21	CAP	Written capitalised	4.585
22	IRREG	Irregular plural	23.111
23	WORD	The actual word	150.837
24	PHON	Phonetic transcription	38.420
25	DPHON	Edited phonetic transcription	136.982
26	STRESS	Stress pattern	38.390

Fonte: Adaptado de Wilson (1988)

após a separação em frases têm-se:

$$Frases = [\text{ 'Exemplo de texto simples.', ' Segunda frase.' }] \tag{2.4}$$

e a separação de palavras da primeira frase resulta em:

$$Palavras = [\text{ 'Exemplo', 'de', 'texto', 'simples', '.' }]. \tag{2.5}$$

Dentro do processo de normalização também são realizadas a eliminação de espaços e salto de linhas, bem como caracteres inválidos. A conversão dos caracteres das palavras para minúsculas (*lowercase*) também pode ser realizado, fazendo com que todas as palavras estejam em um mesmo padrão. Geralmente nesta etapa é realizada a retirada dos caracteres de formatação, tais como os *html tags*, e conversão de conjuntos de caracteres (*charset*), quando os textos são oriundos de plataformas ou idiomas distintos.

$$Normalizado = [\text{ 'exemplo', 'de', 'texto', 'simples', '.' }] \tag{2.6}$$

2.6.5.2 *Remoção de Stop Words*

A remoção das chamadas *Stop Words* consiste em eliminar do texto as palavras que aparecem com frequência e que tem pouco agregado sintático. Tendo como exemplo as palavras de (2.5), após o processo de remoção obtêm-se:

$$SemStopWords = [\ \text{'exemplo'}, \text{'texto'}, \text{'simples'}, \text{'.'}\ ]. \tag{2.7}$$

2.6.5.3 *Remoção de Valores Numéricos*

Tal qual as *Stop Words*, os valores numéricos, bem como as unidades, tal qual "kg", "mm", "milhão", não agregam valor sintático e podem ser excluídas.

2.6.5.4 *Correção Ortográfica*

Em função da origem dos dados, pode ser interessante a realização de um processo de correção ortográfica, a fim de substituir as palavras que possuem pequenos erros de digitação por seu significado sintático correto. Isto deve ser uma preocupação principalmente quando a origem dos dados obtidos de forma *online*, onde os autores estão sujeitos a pequenos erros não intencionais e normalmente o texto não é revisado.

2.6.5.5 *Radicalização*

O processo de radicalização (*stemming*) consiste em reduzir as palavras para seu radical, sendo que os verbos vão para a forma infinitiva. As palavras "aluno", "aluna" e "alunas" são reduzidas para "alun", assim como os verbos "corriam", "correm" e "correu" são transformados para "correr". Desta forma a diversificação de palavras do *corpus* é diminuída, facilitando o processamento.

2.6.5.6 *Categorização*

O processo de categorização (*part-of-speech tagger*) consiste em associar uma categoria sintática a cada uma das palavras do texto. Continuando com o exemplo (2.5), após a categorização, o conjunto de dados obtidos seria:

$$Categorias = [\ (\text{'exemplo'}, \text{'NOUN'}), (\text{'texto'}, \text{'NOUN'}), (\text{simples}, \text{'ADJ'}), (\text{'.'}, \text{'.'})\ ]. \tag{2.8}$$

onde os rótulos "NOUN", "ADJ" e "." indicam as categorias de substantivos, adjetivos e pontuação, respectivamente. O processo de radicalização não consiste somente na associação de uma palavra com a sua categoria, mas precisa levar em consideração o contexto da palavra dentro da frase. Por exemplo a palavra "olhar" pode ser classificada como "VERB" na frase "Vou olhar para o relógio" ou como "NOUN" na frase "Você tem um olhar crítico!".

2.6.6 *Bag of Words*

O modelo *Bag of Words* é uma forma de representar o texto como vetor contendo as diversas palavras encontradas neste texto sendo associadas a informações referentes à frequência de ocorrência destas, sem levar em consideração a gramática ou ordem de ocorrências das palavras no texto.

A transformação de um conjunto de palavras em vetores numéricos permite o processamento dos textos em algoritmos de aprendizado de máquina e arquiteturas *deep learning*. Algumas técnicas foram desenvolvidas, utilizando como referência a frequência de ocorrência dos termos dentro do texto. As mais utilizadas são:

- Vetor de Contagem;
- *Term Frequency–Inverse Document Frequency (TF-IDF)*;
- Vetor de Coocorrência.

2.6.6.1 Modelo de Espaço Vetorial

O Modelo de Espaço Vetorial é uma representação de conjuntos de textos, ou documentos, na forma de vetores de identificadores (Salton et al., 1975). Esta forma de representação facilita a aplicação de processos computacionais, como por exemplo, para processos de classificação de texto.

Considerando o conjunto de documentos D_1 até D_n, conforme ilustrado na Figura 2.6, pode-se dizer que cada documento D_i pode ser representado como

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{in}) \tag{2.9}$$

onde d_{ij} representa o peso do termo j no documento i. Este peso pode ser atribuído de acordo com a importância de cada termo no documento, ou simplesmente por um valor como "1" ou "0", correspondendo à presença ou ausência deste termo no documento, respectivamente.

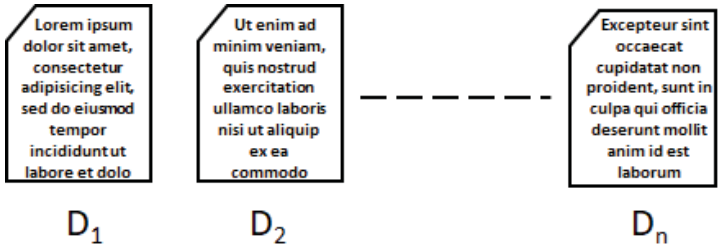


Figura 2.6: Conjunto de Documentos

A Figura 2.7 apresenta um conjunto de 3 documentos, cada qual contendo 3 termos, representados em um espaço tridimensional. Neste caso, a similaridade entre dois documentos pode ser verificada pela distância vetorial entre estes. Assim sendo, dois documentos que fossem idênticos, teriam um ângulo nulo entre eles.

2.6.6.2 Modelo nGRAM

Em termos de probabilidade e linguística computacional, o termo "*n-gram*", que na sequencia deste documento será referenciado como nGRAM, corresponde a uma sequencia de "n" itens componentes de um texto fornecido. Em função da aplicação utilizada, estes itens podem ser palavras, fonemas, sílabas ou caracteres. Utiliza-se as denominações de "*unigram*", "*bigram*" e "*trigram*" quando os valores de "n" correspondem a um, dois ou três, respectivamente. O modelo nGRAM é um tipo de modelo de linguagem para a predição de um próximo elemento em uma sequência na forma de uma ordem $(n-1)$, no modelo probabilístico de Markov (Gagniuc, 2017). Na área de identificação e classificação automática de documentos, a utilização das

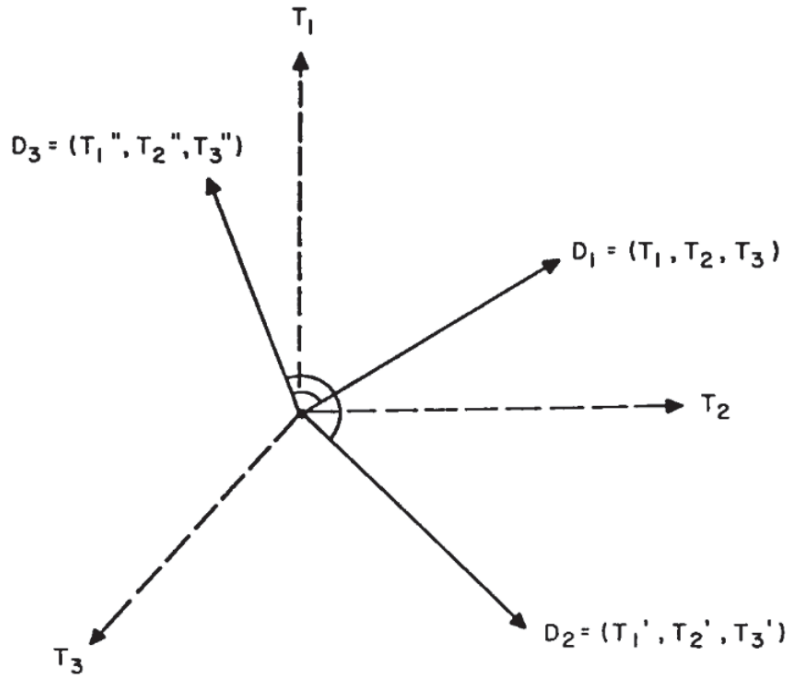


Figura 2.7: Representação Vetorial de Documentos

Fonte: Salton et al. (1975)

informações estatísticas das ocorrências de nGRAM produz relevantes ganhos neste processo (Cavnar et al., 1994). Considerando a pequena frase de exemplo, "a bela e a fera", a geração das sequências tem como resultado os seguintes conjuntos:

***unigram*:** ('a'), ('bela'), ('e'), ('a'), ('fera')

***bigram*:** ('a', 'bela'), ('bela', 'e'), ('e', 'a'), ('a', 'fera')

***trigram*:** ('a', 'bela', 'e'), ('bela', 'e', 'a'), ('e', 'a', 'fera') .

Após o processo de obtenção de um conjunto de nGRAM oriundos do texto, pode ser realizado um levantamento estatístico da frequência de ocorrência destes, gerando um conjunto de características que represente o texto.

2.6.6.3 Vetor de Contagem

A técnica de "Vetor de Contagem" consiste na obtenção de uma matriz que represente a frequência de ocorrência de cada palavra em cada documento de um *corpus*.

Considerando como exemplo um *corpus* hipotético D contendo dois documentos d_1 e d_2, com o seguinte conteúdo:

$$D = [d_1, d_2] \tag{2.10}$$

$$d_1 = [\text{'a', 'casa' , 'tem', 'a', 'janela', 'da', 'cor', 'azul'}] \tag{2.11}$$

$$d_2 = [\text{'a', 'porta', 'da', 'casa', 'abriu'}]. \tag{2.12}$$

Considerando W o conjunto único das palavras encontradas em D

$$W = [w_1, w_2, ...,w_n] \ \ |w_i \in D \tag{2.13}$$

$$W = [\text{'a', 'casa', 'tem', 'janela', 'da', 'cor', 'azul', 'porta', 'abriu'}]. \tag{2.14}$$

No caso da palavra 'a', esta aparece duas vezes em d_1 e uma vez em d_2. Seguindo com este raciocínio para todas as palavras do *corpus*, é realizada a montagem da matriz M_{cv} que representa este *corpus*, de acordo com a técnica *Count Vector*.

	a	*casa*	*tem*	*janela*	*da*	*cor*	*azul*	*porta*	*abriu*
d_1	2	1	1	1	1	1	1	0	0
d_2	1	1	0	0	0	0	0	1	1

$$M_{cv} = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \tag{2.15}$$

Cada coluna k da matriz M_{cv} representa o vetor da frequência de ocorrência de cada palavra $w_k \ \ |w_k \in W$, dos documentos d_i do *corpus* D.

2.6.6.4 Term Frequency–Inverse Document Frequency (TF-IDF)

O valor TF-IDF, que significa frequência do termo inverso da frequência no documento (*term frequency–inverse document frequency*), é uma medida estatística que tem como objetivo indicar a relevância de uma palavra em um documento ou em um *corpus* (Ullman, 2011). O valor TF-IDF é proporcional ao número de vezes que uma palavra ocorre em um documento, mas também é ajustado em função da frequência da palavra no corpus, equilibrando a medida em função das palavras que possuem mais ocorrências. Considerando que $f_{w,d}$ representa a frequência de ocorrência de uma palavra w em um documento d, obtêm-se a expressão da frequência do termo (TF).

$$\text{tf}(w,d) = \frac{f_{w,d}}{\sum\limits_{w_k \in d} f_{w_k,d}} \tag{2.16}$$

O inverso da frequência no documento (IDF) é uma medida que representa o quanto uma palavra é comum ou rara em um documento, ou seja, o quanto ela é importante para representar um documento dentro de um corpus D. É resultante da divisão do número total de documentos N existentes no *corpus* D pelo número de documentos $|\{d \in D : w \in d\}|$ em que esta palavra ocorre.

$$\text{idf}(w,D) = log\frac{N}{|\{d \in D : w \in d\}|} \tag{2.17}$$

Têm-se então, finalmente, a obtenção do TF-IDF.

$$\text{tfidf}(w,d,D) = \text{tf}(w,d) \times \text{idf}(w,D) \tag{2.18}$$

Considerando como exemplo um *corpus* hipotético D contendo dois documentos d_1 e d_2 com o seguinte conteúdo:

$$D = [d_1, d_2] \tag{2.19}$$

$$d_1 = [\text{'a', 'casa' , 'tem', 'a', 'janela', 'da', 'cor', 'azul'}] \tag{2.20}$$

$$d_2 = [\text{'a', 'porta', 'da', 'casa', 'abriu'}]. \tag{2.21}$$

No caso do valor TF para a palavra 'casa', pode ser observado que ela ocorre nos dois documentos, mas o documento d_2 tem mais palavras.

$$\text{tf('casa', } d_1) = \frac{1}{7} \approx 0{,}1428 \tag{2.22}$$

$$\text{tf('casa', } d_2) = \frac{1}{5} = 0{,}2 \tag{2.23}$$

Já para o IDF neste mesmo caso, temos que a palavra 'casa' aparece nos dois documentos.

$$\text{idf('casa,}D) = log\left(\frac{2}{2}\right) = 0 \tag{2.24}$$

Com estes valores parciais, pode ser obtido o TF-IDF.

$$\text{tfidf('casa', } d_1) = 0{,}1428 \times 0 = 0 \tag{2.25}$$

$$\text{tfidf('casa', } d_2) = 0{,}2 \times 0 = 0 \tag{2.26}$$

O que indica que a palavra 'casa' neste contexto não é uma característica relevante, visto que ocorre em todos os documentos do *corpus*.

Já no caso da palavra 'porta', que ocorre em somente um dos documentos, o resultado é bem diferente.

$$\text{tf('porta', } d_1) = \frac{0}{7} = 0 \tag{2.27}$$

$$\text{tf('porta', } d_2) = \frac{1}{5} = 0{,}2 \tag{2.28}$$

$$\text{idf('porta,}D) = log\left(\frac{2}{1}\right) \approx 0{,}301 \tag{2.29}$$

$$\text{tfidf('porta', } d_1) = 0 \times 0{,}301 = 0 \tag{2.30}$$

$$\text{tfidf('porta', } d_2) = 0{,}2 \times 0{,}301 \approx 0{,}0602 \tag{2.31}$$

2.6.6.5 Vetor de Coocorrência

A técnica de "Vetor de Coocorrência" está baseada na ideia de que as palavas similares tendem a ocorrer próximas em um texto. A "coocorrência" de um par de palavras é o número de vezes que estas ocorrem em uma determinada janela de contexto. A "janela de contexto" por sua vez é o espaço compreendido por "n" palavras à direita e à esquerda de uma determinada palavra.

Considerando como exemplo um documento hipotético d com o seguinte conteúdo:

$$d = [\text{'o', 'doce', 'de', 'batata', 'doce', 'é', 'o', 'doce', 'mais', 'doce' }]. \tag{2.32}$$

Considerando W o conjunto único das palavras encontradas em d.

$$W = [w_1, w_2, ...,w_n] \ \ |w_i \in d \tag{2.33}$$

$$W = [\text{'o', 'doce', 'de', 'batata', 'é', 'mais'}] \tag{2.34}$$

Elaborando uma matriz do coocorrência M, quadrada, de tamanho n, onde $n = |W|$. Cada elemento $m_{i,j}$ contém a coocorrência entre as palavras w_i e w_k, considerando uma janela de contexto de tamanho "2".

	o	*doce*	*de*	*batata*	é	*mais*
o	0	3	1	0	1	1
doce	3	1	2	2	2	2
de	1	2	0	1	0	0
batata	0	2	1	0	1	0
é	1	2	0	1	0	0
mais	0	2	0	0	0	0

$$M = \begin{pmatrix} 0 & 3 & 1 & 0 & 1 & 1 \\ 3 & 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 0 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 & 1 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{2.35}$$

2.6.7 *Word Embedding*

O termo "incorporação de palavra", mais conhecido como *Word Embedding* (WE), refere-se a um conjunto de técnicas onde as palavras ou textos de um *corpus* são associadas a vetores de números reais. Ou seja, corresponde a incorporação de um espaço com uma dimensão por palavra para um espaço vetorial contínuo com muito menos dimensões.

Diversos são os métodos utilizados para realizar esta associação, sendo que os mais comuns são:

- redes neurais;
- redução de dimensionalidade com matriz de co-ocorrência;
- modelos probabilísticos;
- base de conhecimento e
- representação explícita nos termos do contexto.

2.6.7.1 *Modelos Word2Vec*

Os modelos de predição *Word2Vec* são o resultado de uma iniciativa do Google, por meio de um grupo liderado por Tomas Mikolov. Estes modelos utilizam o contexto linguístico das palavras, por meio de redes neurais superficiais (*Shallow Neural Networks*) (SNN) treinadas para esta predição (Mikolov et al., 2013). O termo SNN refere-se ao tipo de rede neural (NN) que somente possui uma camada oculta, ao contrário das DNN (*Deep Neural Network)* que possuem diversas camadas ocultas.

A partir de um conjunto de textos fornecidos, estes modelos produzem um espaço vetorial, com centenas de dimensões, associando para cada termo do *corpus* um vetor numérico correspondente. Os vetores são criados de forma que os termos que compartilham contextos comuns são posicionados próximos, no espaço vetorial. Quando (Mikolov et al., 2013) propuseram o *Word2Vec*, também especificaram os modelos *Skip-gram* e *Continuous Bag of Words* (CBOW), ilustrados na Figura 2.8. O modelo CBOW é utilizado para realizar a predição de uma palavra dentro de um contexto fornecido. Por sua vez o modelo *Skip – Gram* é utilizado para realizar a predição de um contexto a partir de uma palavra fornecida.

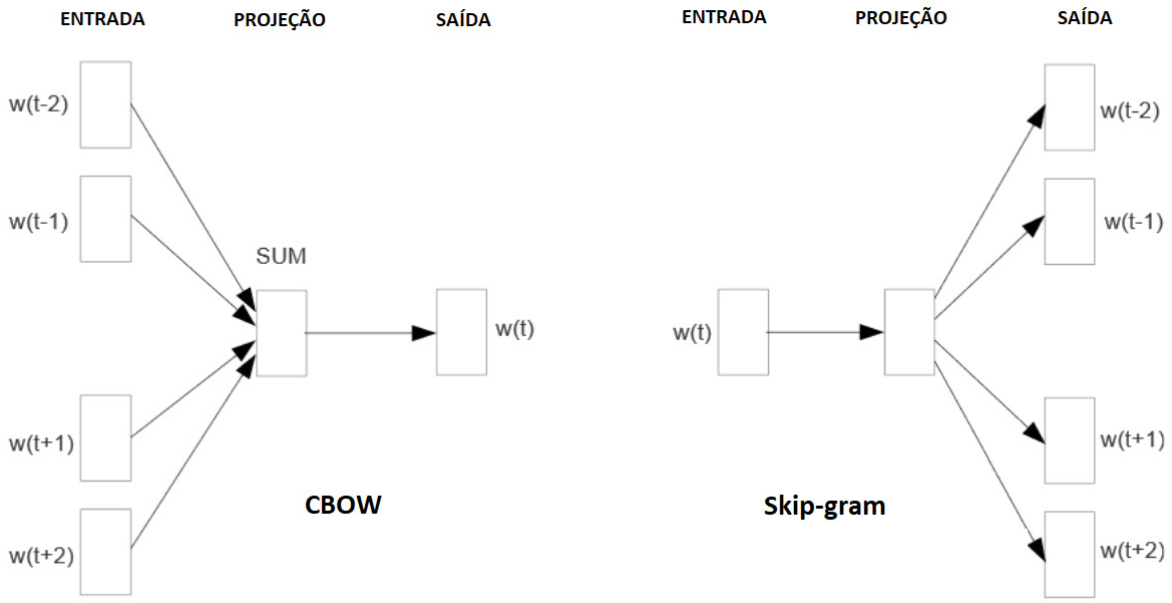


Figura 2.8: Modelos CBOW e *Skip-gram*

Fonte: Adaptado de Mikolov et al. (2013)

2.6.7.2 *Gensim*

O *Gensim* é um conjunto de ferramentas de código livre, desenvolvido inicialmente por Řehůřek e Sojka (2011), com o objetivo de realizar modelagens no vetor espacial, incluindo os algoritmos *Word2Vec* e *Doc2vec* na linguagem *Python*, utilizando as bibliotecas *NumPy* e *SciPy*. O Algoritmo *Doc2Vec* é uma extensão do *Word2Vec* voltado ao tratamento de documentos do tipo texto.

2.6.7.3 Deeplearning4j

O *Deeplearning4j* consiste em uma biblioteca para aprendizado de máquina desenvolvida em JAVA, que também permite a execução dos algoritmos *Word2Vec* e *Doc2vec*, desenvolvida pela equipe liderada por Adam Gibson (Patterson e Gibson, 2017).

2.6.7.4 fastText

O *fastText* é uma biblioteca desenvolvida pelo Centro de Pesquisa para Inteligência Artificial do *Facebook* (FAIR) para o aprendizado de *word embeddings* (Joulin et al., 2016). Em setembro de 2017 *Facebook* disponibilizou esta ferramenta com bases treinadas em 294 linguagens.

2.7 Classificação

Esta seção apresenta os conceitos relacionados às técnicas de classificação dos dados obtidos dos textos com o objetivo de identificação de perfil de personalidade. Inicialmente são apresentados os conceitos de classificação dentro da área de Aprendizado de Máquina, discutidas os modelos e técnicas utilizados para realizar o comparativo paramétrico de classificadores, as ferramentas utilizadas para estas atividades, finalizando com uma descrição dos classificadores mais utilizados nos experimentos encontrados na literatura e na presente pesquisa.

2.7.1 Aprendizado de Máquina

A área de Aprendizado de Máquina (*Machine Learning*), que doravante será referenciado como ML, pode ser atribuída como um campo da Computação que utiliza técnicas estatísticas para propiciar aos computadores a habilidade de aprender baseado em dados fornecidos. O termo "aprender" está associado a diversos conceitos, tais quais: adquirir conhecimento de algo pelo estudo ou experiência; obter informações pela obervação; fixar na memória; ser informado ou verificar e receber instruções. Em termos computacionais o significado de aprendizado está relacionado à mudança de comportamento de maneira que propicie um melhor desempenho no futuro. Ou seja, o conceito de aprendizado está relacionado à melhoria de performance de um processo computacional (Witten et al., 2016).

O desenvolvimento e utilização de técnicas computacionais que propiciem o aprendizado e construção de sistemas que permitam a aquisição de conhecimento de forma automática, são objetivos da área de ML. O processo de aprendizagem pode ser obtido por processos computacionais que tomem decisões baseado em experiências anteriores que foram bem sucedidas. Ao passo que os humanos estão habituados à utilização do raciocínio dedutivo para obter uma nova informação, baseado no relacionamento lógico de outras informações adquiridas anteriormente, os processos computacionais utilizam o método indutivo, em que a obtenção de novas informações é baseada na inferência lógica realizada em um conjunto de exemplos.

Em processos computacionais de aprendizado, usualmente trabalha-se com um conjunto de dados previamente conhecido, chamado de "base de treinamento", e um outro conjunto de dados para os quais queira se fazer a inferência, chamada "base de teste". Cada "elemento" das bases de treinamento e teste é composto por um conjunto de atributos, chamados de "características", que representam este elemento em termos quantitativos. Cada elemento também pode estar associado a uma "classe" (ou a mais de uma). Um exemplo de base de treinamento que pode ser utilizada em um processo de classificação, conforme apresentado na Figura 2.9.

Elemento	**Características**			**Classe**
ID	Altura	Peso	Idade	Tipo
1	1,32	32,3	12	0
2	1,75	64,2	27	1
3	1,64	65,2	45	0
:	:	:	:	:
50	1,56	45,1	19	0

Figura 2.9: Exemplo de Base de Treinamento

Os métodos de aprendizado podem ser classificados em:

Supervisionado - ocorre quando os elementos da base de treinamento estão associados a classes previamente definidas. Ou seja, a base de treinamento foi previamente classificada, geralmente por especialistas humanos. Pode ser ainda divida em:

Classificação - quando as classes são representadas por valores discretos;

Regressão - quando as classes são representadas por valores contínuos;

Não Supervisionado - neste caso não existe uma classificação prévia dos elementos, sendo que um conjunto de observações deve ser realizado com o objetivo de separar os elementos em classes (*clusters*).

Em um processo de classificação em que somente duas classes estão presentes é denominado de "Classificação Binária", como o exemplo ilustrado na Figura 2.9. Neste exemplo, cada elemento do conjunto é associado a uma classe "0" ou "1". Quando mais de duas classes são utilizadas, o processo é denominado "Classificação Multiclasses". Um exemplo deste caso é quando pretende-se separar um conjunto de animais vertebrados nas categorias "mamífero", "ave", "réptil", "peixe" e "anfíbio", baseado em um conjunto de características fornecidos para cada elemento. Nos casos em que um determinado elemento está associado a mais de uma classe, o processo é denominado "Classificação Multirótulo". Um exemplo deste tipo é quando artigos de um jornal estão associados a diversas categorias como "esporte", "lazer", "internacional", sendo que cada elemento pode pertencer a mais de uma categoria simultaneamente.

Um "Classificador" pode ser descrito como um algoritmo utilizado para realizar a identificação automática da classe a que pertencem cada um dos elementos da base de teste, em um processo de classificação. Também associado à função ou processo computacional que realiza esta atividades. Existe uma grande gama de classificadores disponíveis, sendo que alguns métodos de classificação, relacionados ao tema da pesquisa, são tratados nesta seção.

2.7.2 Comparação de Classificadores

O processo de comparação de classificadores está associado à necessidade de comparar dois métodos diferentes de classificação aplicados a um mesmo problema. Isto pode ser feito estimando o erro obtido submetendo os dois classificadores a uma bateria de testes sucessivas, utilizando as mesmas bases de treinamento e teste, escolhendo o classificador que apresentar o menor erro. Este é um problema padrão para pesquisadores da área de aprendizado de máquina. Se um novo algoritmo é proposto, este deve demonstrar que a melhoria observada não é uma simples obra do acaso (Witten et al., 2016).

A performance de um classificador pode ser medida em termos da "taxa de erro". O classificador realiza a estimativa, ou predição, a que classe pertence um elemento. Tendo o conhecimento prévio a que classe pertence este elemento, resultante de uma classificação manual prévia por exemplo, é verificado se o classificador estimou a classe correta. Neste caso a predição é registrada como "sucesso", e caso contrário, é registrada como "erro". Considerando uma base de teste contendo n elementos, na qual foram registrados n_s predições corretas (sucesso) e n_e predições incorretas (erro), sendo $n = n_s + n_e$, pode-se dizer que taxa de erro tx_{erro} de um classificador pode ser expressa como:

$$tx_{erro} = \frac{n_e}{n_e + n_s}. \tag{2.36}$$

No caso específico de classificação binária, como a ilustrada na Figura 2.9, as classes a que pertencem cada elemento podem ser denominadas de "positivas" e "negativas". Neste exemplo podemos associar a classe positiva ao rótulo "1" e a classe negativa ao rótulo "0".

Realizando um processo de classificação, e comparando os resultados retornados pelo classificador com os valores reais de cada elemento, podem ocorrer quatro situações distintas, que são:

VP (Verdadeiro Positivo) - o valor predito é positivo e o valor real é positivo;

VN (Verdadeiro Negativo) - o valor predito é negativo e o valor real é negativo;

FP (Falso Positivo) - o valor predito é positivo e o valor real é negativo e

FN (False Negativo) - o valor predito é negativo e o valor real é positivo.

Estes valores podem ser aplicados em uma tabela, denominada "Tabela de Confusão", ou como mais comumente é chamada, de "Matriz de Confusão" (*Confusion Matrix*), conforme ilustrado na Figura 2.10.

		Predição	
		0	1
Real	0	**V**erdadeiro **N**egativo	**F**also **P**ositivo
	1	**F**also **N**egativo	**V**erdadeiro **P**ositivo

Figura 2.10: Matriz de Confusão

A Matriz de Confusão permite a obtenção de valores mais apurados sobre a performance de um classificador. Por exemplo, a obtenção da taxa de erro em uma base desbalanceada, na qual o número de elementos de classe positiva não é igual ao número de classes negativas, pode levar a avaliações incorretas sobre a performance de um classificador.

2.7.2.1 Acurácia

A Acurácia representa a proporção de predições corretas em relação ao total de elementos. Esta medida pode ser utilizada para avaliar um conjunto de classificadores, ou seja, quanto maior a acurácia apresentada por um classificador, maior será a taxa de acertos deste classificador.

$$acc = \frac{TP + TN}{VP + VN + FP + FN} \tag{2.37}$$

2.7.3 Validação Cruzada

A técnica de validação cruzada é utilizada para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados fornecido. Consiste basicamente na separação de um conjunto de dados que será utilizado para avaliar a performance de um classificador, em conjuntos menores, mutuamente exclusivos. Estes subconjuntos são agrupados, de acordo com um critério estabelecido, em dois grupos, formando uma base de treinamento e uma base de testes (Kohavi et al., 1995). É recomendável que durante a divisão do conjunto de dados, a proporção de cada uma das classes existentes na base original seja preservada nos conjuntos menores, em um processo conhecido como estratificação (Refaeilzadeh et al., 2009).

As técnicas de validação cruzada podem ser "exaustivas" e "não exaustivas". Os métodos exaustivos são aqueles que testam todas as possíveis maneiras de separar o conjunto de dados, ao passo que os métodos não exaustivos realizam uma separação aleatória das bases. Dentre as modalidades de técnicas não exaustivas, o método "*k-fold cross validation*" é o mais utilizado. Neste método, a base de dados original é separada em k subconjuntos de tamanho menor, mutuamente excludentes, e que tendam a ter tamanhos iguais. Dos "k" subconjuntos, um é utilizado como base de testes e os restantes $k-1$ subconjuntos são agrupados e utilizados como base de treinamento. O processo de validação cruzada é repetido k vezes, utilizando cada subconjunto como base de testes. O valor médio dos k resultados obtidos é obtido para produzir um resultado único dos testes realizados com o classificador. A Figura 2.11 ilustra o processo de validação cruzada *k-fold* onde $k = 3$.

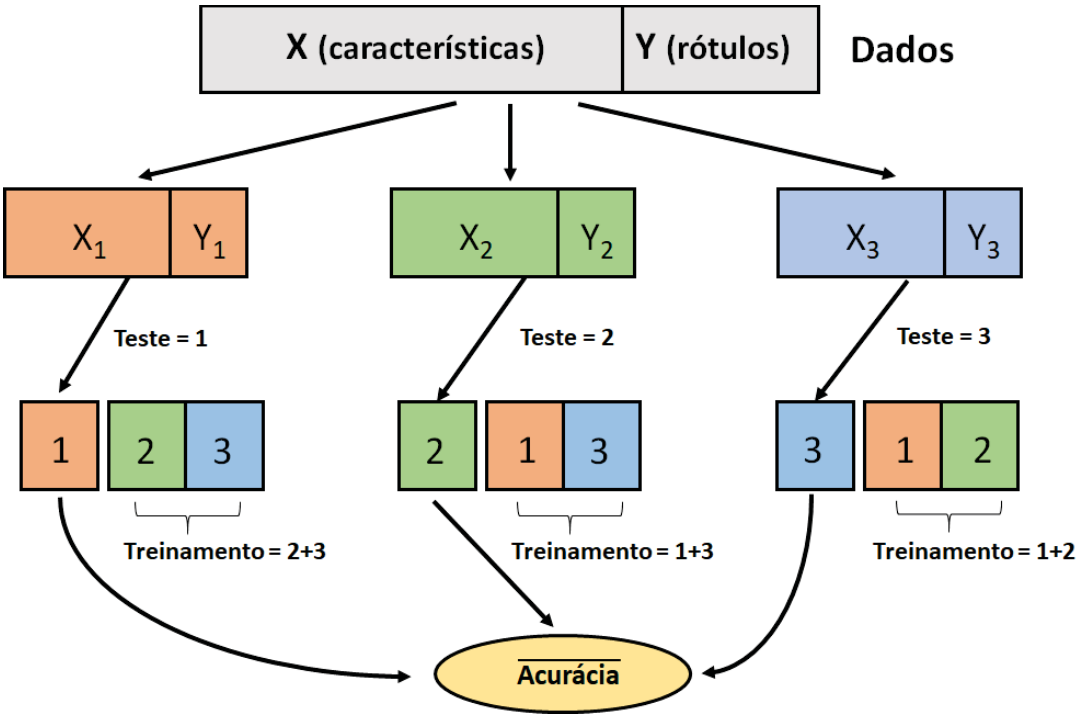


Figura 2.11: Validação Cruzada *3-fold*

Quando este tipo de validação é utilizada, a acurácia do classificador é apresentada como uma acurácia média, resultante das k combinações e classificações realizadas.

$$\overline{acc} = \frac{\sum_{i=1}^{k} acc_i}{k} \tag{2.38}$$

Além da acurácia média, deve ser apresentado o desvio padrão σ, obtido a partir dos valores individuais acc_i de cada uma das k classificações intermediárias.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(\overline{acc} - acc_i)^2}{k}} \tag{2.39}$$

2.7.4 Ferramentas

A realização dos testes comparativos entre classificadores pode ser realizada de uma forma manual, onde os ensaios são realizados com os algoritmos de classificação desenvolvidos em alguma linguagem de programação, com ou sem o uso de bibliotecas (API), ou por meio de uma ferramenta de *software* que realize os ensaios.

Dentre as ferramentas de software mais utilizadas para esta finalidade têm-se o WEKA (*Waikato Environment for Knowledge Analysis*) (Garner, 1995; Hall et al., 2009), que teve suas origens na Universidade de Waikato, Nova Zelândia, em 2006. Desenvolvida em Java, sendo multiplataforma, esta ferramenta tem sido utilizada amplamente pela comunidade acadêmica e científica para realização de testes na área de aprendizado de máquina, visto agregar um grande número de algoritmos de classificação oriundos de diversas abordagens e paradigmas. A Figura 2.12 ilustra a tela de entrada desta ferramenta.

Figura 2.12: Ferramenta WEKA

Uma das bibliotecas mais utilizadas para a realização de testes de classificadores utilizando programação em linguagem de alto nível, é o *SCIKIT-LEARN* (Pedregosa et al., 2011; Buitinck et al., 2013), uma extensa biblioteca de algoritmos para aprendizado de máquina de código aberto, desenvolvida inicalmente por David Cournapeau em 2007 na linguagem *Python* (Van Rossum e Drake, 2003). Atualmente um grande grupo de desenvolvedores colabora na manutenção e ampliação desta biblioteca. Os experimentos realizados no escopo da presente pesquisa foram obtidos com a utilização de programas em *Python* utilizando o *sci-kit learn*.

2.7.5 Classificador *k-Nearest Neighbors*

O algoritmo *k-Nearest Neighbors* (kNN) é um modelo não paramétrico que tem como objetivo identificar a que classe pertence um determinado elemento, baseado na distância que este elemento tem dos demais elementos da base de treinamento. A identificação é realizada verificando a classe a que pertencem os "k" vizinhos mais próximos, sendo o valor de "k" arbitrado para cada caso, sendo indicado a utilização de números ímpares. O processo de escolha para qual classe será atribuído o elemento que se esteja estimando, é baseado no valor majoritário das classes associadas aos "k" vizinhos mais próximos.

A Figura 2.13 ilustra o caso de uma série de dados com os elementos da base de treinamento previamente categorizadas como classes A e B. O problema é estimar a qual classe pertence o elemento da base de teste (em azul), baseado na distância d_i de cada um dos seus vizinhos mais próximos. Neste exemplo foi adotado $k = 5$, sendo que na área que delimita os 5 vizinhos mais próximos, ilustrada em tracejado, a votação majoritária indica que o novo elemento pertence a classe A. Em função da distribuição dos elementos da base, a utilização de valores diferentes de k pode resultar em estimativas para classes distintas.

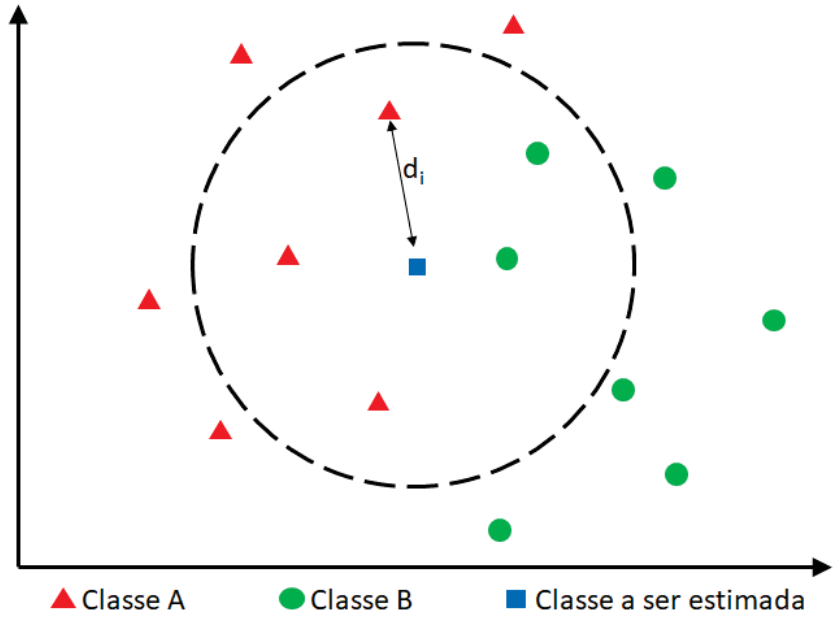


Figura 2.13: Vizinhos Mais Próximos

Este classificador é um dos mais simples de serem codificados apresentando também baixo custo de processamento. O cálculo da distância pode ser feito utilizando diversas abordagens, sendo que a mais simples é a adoção do cálculo da Distância Euclidiana dos pontos, considerando um espaço multidimensional com tamanho igual ao número de características da base. Considerando os pontos n-dimensionais $A = (a_1, a_2, \ldots, a_n)$ e $B = (b_1, b_1, \ldots, b_n)$, a distância euclidiana d pode ser expressa como:

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2}. \tag{2.40}$$

2.7.6 Classificadores Naïve Bayes

Este termo refere-se a um conjunto de classificadores probabilísticos baseados na aplicação do teorema de Bayes e assumindo uma forte independência entre as características. Um classificador probabilístico, ou estatístico, é aquele que consegue prever a distribuição probabilística de um conjunto de classes. Considerando um conjunto finito X de elementos, sendo que cada elemento x_i está associado a uma classe y_i, pertencente a um conjunto finito de classes Y, pode-se dizer que a classe estimada $\hat{y}_i$ de um elemento genérico x_i pode ser obtido como

$$\hat{y}_i = f(x_i) \tag{2.41}$$

onde $f(x)$ representa uma função que associa uma classe a cada elemento do conjunto X.

No caso dos classificadores probabilístico, esta função $f(x)$ é expressa como distribuições condicionais $P(Y|X)$, que demonstram a probabilidade de um determinado elemento do conjunto estar associado a cada uma das classes. Considerando que para um elemento $x_i \in X$, são associadas probabilidades de ocorrência para todos os elementos $y \in Y$, com a utilização da regra de decisão ótima (Bishop, 2006) pode-se expressar que uma classe $\hat{y}_i$ é determinada pela classe com maior probabilidade de ocorrência.

$$\hat{y}_i = \text{argmax}_y \, P(Y = y|X) \tag{2.42}$$

O teorema de Bayes descreve a probabilidade da ocorrência de um evento, baseado nas condições prévias que podem estar relacionadas a este evento (Uspensky, 1937). Considerando A e B dois eventos têm-se que este teorema pode ser descrito como

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.43}$$

sendo que $P(B) \neq 0$.

Os classificadores mais utilizados nesta categoria são o *Gaussian Naïve Bayes* (GNB), o *Multinomial Naïve Bayes* (MNB) e o *Bernoulli Naïve Bayes* (BNB).

2.7.7 Classificador *Logistic Regression*

Este tipo de algoritmo está baseado na técnica de Regressão Logística (LR), que é um recurso que permite estimar a probabilidade associada à ocorrência de determinado evento baseado em uma ou mais variáveis de predição, e foi desenvolvida por Cox (1958). A LR pode ser binomial, ordinária ou multinomial. A binomial, ou binária, trata das situações em que saída observada para a variável dependente pode ter somente dois valores, como nos casos de classificação binária. A multinomial engloba os casos em que os valores de saída assumem mais de dois valores e a ordinária trata das situações em que as variáveis dependentes estão ordenadas. Nesta seção somente será abordado o caso da regressão binária (Nasrabadi, 2007).

As variáveis de predição podem ser contínuas ou categóricas. Ao contrário da regressão linear, a regressão logística utiliza varáveis de predição dependentes que pertencem a um número limitado de categorias. Este tipo de predição utiliza uma função de ligação que associa os valores esperados da resposta aos preditores lineares no modelo, ou seja, a probabilidade de ocorrência dos eventos para cada valor de entrada. Esta pode ser a função de ligação binária $logit$ (Conaway, 1990). A $logit$ de um valor de entrada p que representa uma probabilidade, entre 0 e 1, é expresso como:

$$logit(p) = \ln(\frac{p}{1-p}) = -\ln(\frac{1}{p} - 1). \tag{2.44}$$

A $logit$ do sucesso de ocorrência é ajustado aos preditores. O valor previsto pelo $logit$ é convertido de volta em pares de predição previstos através do inverso do logaritmo natural, ou seja, a função exponencial. Assim, embora a variável dependente observada na regressão logística binária seja uma variável 0 ou 1, a regressão logística estima as chances, como variável contínua, de que a variável dependente seja um sucesso.

2.7.8 Classificador SVM

O conceito de Support Vector Machines (SVM) foi inicialmente apresentado por Cortes e Vapnik (1995) com o objetivo de resolver problemas de classificação binária. A abordagem utilizada por este tipo de algoritmo está baseado em: separação de classes; sobreposição de classes; não linearidade e solução do problema (Meyer e Wien, 2017).

A separação de classes consiste em procurar o hiperplano de separação entre as classes, conforme ilustrado na Figura 2.14, procurando maximizar a margem entre as duas classes. Os

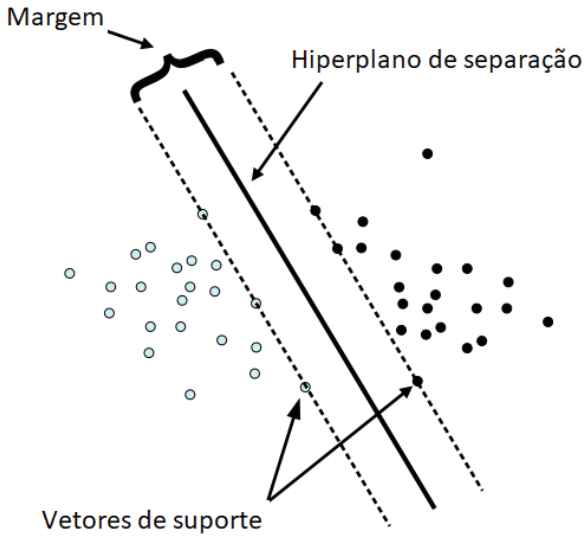


Figura 2.14: SVM - Separação Linear

Fonte: Adaptado de Meyer e Wien (2017)

pontos situados nos limites das margens, formam os vetores de suporte e no meio destes, está o melhor hiperplano de separação.

A sobreposição de classes procura reduzir o peso dos pontos que estão fora da margem discriminante para reduzir a sua influência (*soft margin*). Quando não é possível encontrar um separador linear, os pontos são projetados em um espaço altamente dimensional onde os pontos possam efetivamente ser separados linearmente, através das chamadas técnicas de *kernel*, resolvendo o problema de não linearidade. Para a solução do problema, toda este esforço pode ser formulado como um Problema de Otimização Quadrática (QP) (Floudas e Visweswaran, 1995), que pode ser resolvido utilizando técnicas adequadas.

2.7.8.1 Sequential Minimal Optimization

O *Sequential Minimal Optimization* (SMO) é um algoritmo desenvolvido por Platt (1998) para resolver o problema QP que surgiu durante o treinamento dos classificadores SVM. Este algoritmo tem sido utilizado de forma abrangente para o treinamento de classificadores SVM (Keerthi et al., 2001).

2.7.9 Classificador *Decision Tree*

O modelo preditivo de classificação Árvore de Decisão (*Decision Tree*) utiliza a representação dos elementos de entrada em uma estrutura de árvore, iniciando em um nó (*root*)

que não tem galhos entrantes, ligado a outros nós, que tem somente um galho entrante, em uma estrutura hierárquica crescente. Um nó que tenha galhos saindo é chamado de nó interno ou de teste. Os demais galhos são chamados de folhas ou terminais, ou ainda de nós de decisão. Cada nó interno separa o conjunto de entradas em dois ou mais subconjuntos, de acordo com uma função discreta que é aplicada sobre os atributos de entrada, ou seja, sobre as características do elemento. Para cada folha é associado o valor de uma das classes, que representam o valor de destino mais apropriado, que é a estimativa da classe associada ao elemento em que está sendo feita a predição (Lior et al., 2007).

Foram desenvolvidos muitos algoritmos específicos para realização da árvore de decisão, sendo os mais utilizados:

ID3 - *Iterative Dichotomiser 3*, desenvolvido por Quinlan (1987);

C4.5 - uma extensão do ID3, sendo disponibilizado na plataforma WEKA, com o nome de J48;

CART - *Classification and Regression Trees*;

CHAID - - *Chi-square Automatic Interaction Detection*;

MARS - - *Multivariate adaptive regression splines* voltado a dados numéricos.

2.7.9.1 Classificador Random Forest

O algoritmo *Random Forest* é um tipo de classificador com aprendizado *ensemble*, ou seja, realiza a combinação de diversos classificadores, e através de um processo de votação, indica qual a predição a ser utilizada. O *Random Forest* gera vários classificadores do tipo *Decision Tree*, cada um com suas particularidades e combina o resultado da classificação de todos eles. Com isto os resultados obtidos são superiores ao da utilização de uma *Decision Tree* somente. A Figura 2.15 ilustra o processo de predição para este tipo de classificador.

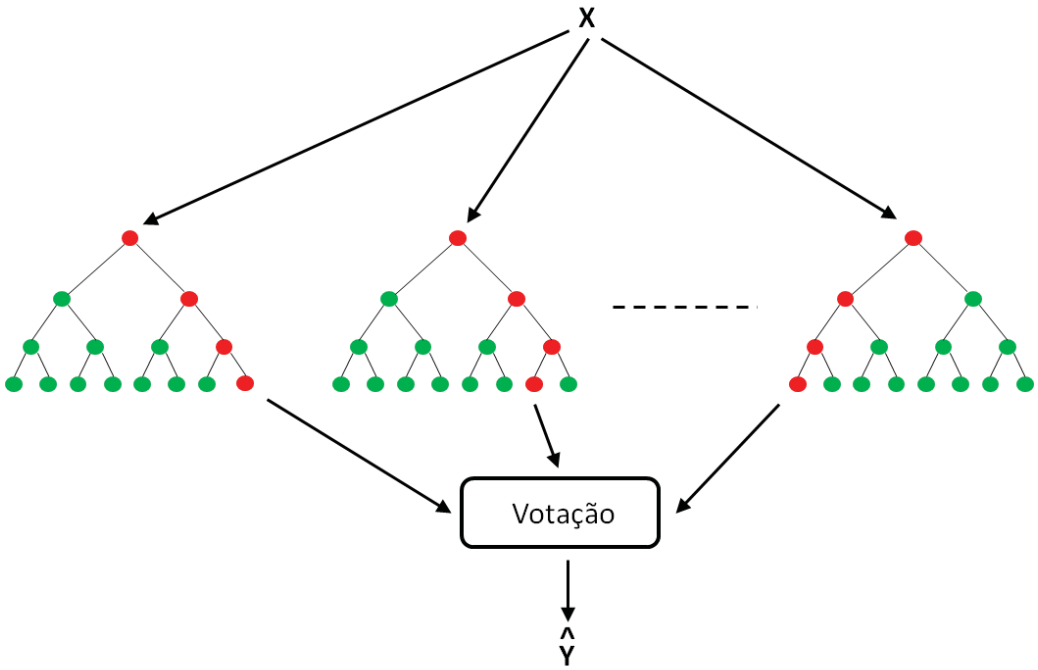


Figura 2.15: *Random Forest*

2.7.10 Classificador *Multilayer Perceptron*

As redes neurais artificias (ANN) são modelos computacionais baseados no comportamento das redes biológicas naturais (NN) existentes nos seres vivos. McCulloch e Pitts (1943) desenvolveram um modelo computacional para ANN denominado *Threshold Logic*, utilizando conceitos matemáticos. Rosenblatt (1958) desenvolveu o modelo *Perceptron*, com o objetivo de modelar como o cérebro humano reconhece padrões visuais e reconhece os objetos. Mas foi nas últimas décadas que ocorreu a utilização em maior escala do conceito de ANN na área de classificação de informações. Este fato deveu-se ao aumento acentuado do poder computacional dos processadores e pela disponibilização de grandes massas de dados para treinamento dos classificadores.

A forma mais básica de uma rede neural é a formada por um *Perceptron*, um neurônio que, a partir dos valores de entrada, realiza a sua soma ponderada, aplica a função de ativação, e envia o resultado para a saída, conforme ilustrado na Figura 2.16.

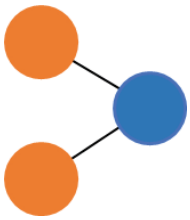


Figura 2.16: Modelo de *Perceptron*

O conceito de rede *Feedforward* é associada a arquitetura em que cada uma das camadas conecta-se a camada seguinte, mas não ocorre um caminho de volta. A rede neural do tipo *Multilayer Perceptron* (MLPC), também conhecida como *Deep Feedforward* (FF) e *Fully Connected* (FC), é uma rede do tipo *Feedforward* que possui diversas camadas ocultas, conforme ilustrado na Figura 2.17. O conceito de *Deep Neural Network* está associado a estas redes com diversas camadas ocultas, que conseguem obter resultados melhores, mesmo com um número reduzido de parâmetros de entrada.

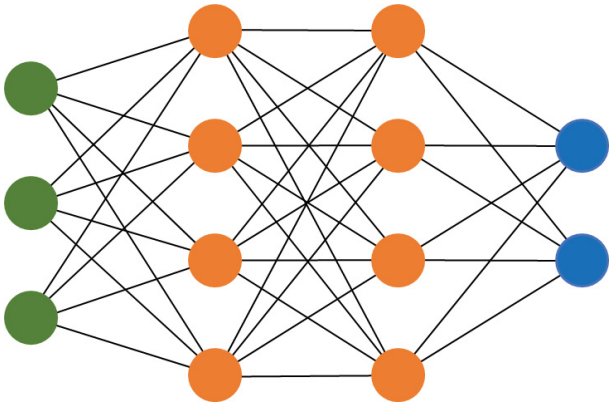


Figura 2.17: Modelo de *Multilayer Perceptron*

2.7.11 Classificador *Ensemble*

O conceito do classificador *Ensemble* está associado a um modelo de predição obtido pela integração de múltiplos classificadores. Em termos gerais, a predição da classe correspondente a

um elemento é realizada por vários classificadores, cada qual obtendo uma estimativa individual. A partir do conjunto de estimativas individuais, é escolhido o valor com ocorrência majoritária, como em um processo de votação, valor este que é indicado como a predição obtida para este elemento.

Em termos gerais o desempenho obtida por esta técnica é superior se comparada com a performance individual de cada um dos classificadores utilizados (Rokach, 2010). A Figura 2.18 ilustra um exemplo deste modelo, utilizando quatro classificadores.

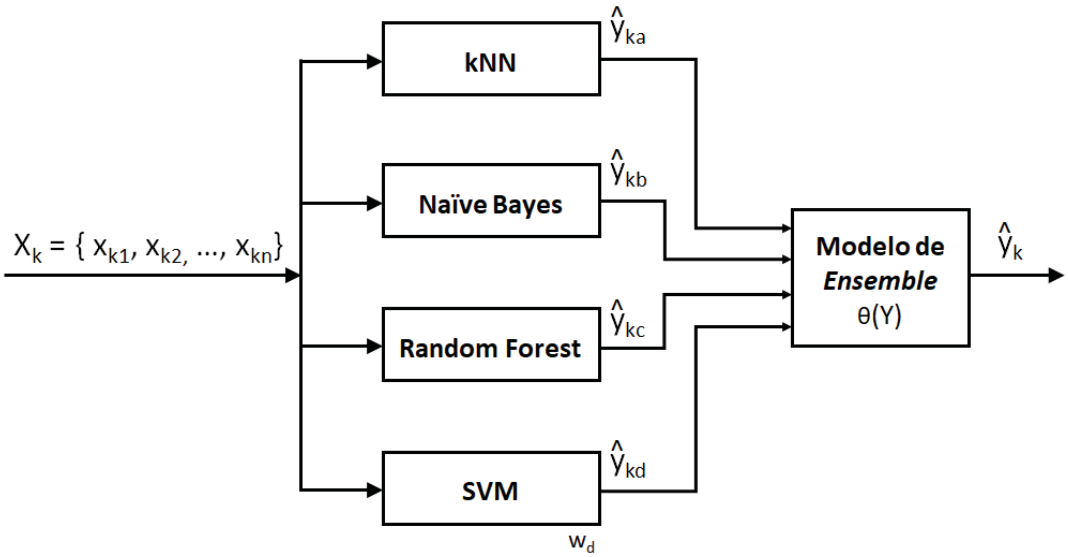


Figura 2.18: Exemplo de Classificador *Ensemble*

Neste exemplo, o valor estimado para um hipotético conjunto de características de entrada, $X_k = \{x_{k_1}, x_{k_2}, \ldots, x_{k_n}\}$, é o valor $\hat{y}_k$. Este valor é obtido pela aplicação de uma função de combinação θ, que é aplicada aos valores estimados por cada classificador individualmente, para a obtenção do valor predito majoritário

$$\hat{y}_k = \theta(\hat{y}_{k_a}, \hat{y}_{k_b}, \hat{y}_{k_c}, \hat{y}_{k_d}). \tag{2.45}$$

A função de combinação θ pode realizar uma combinação simples para a obtenção do voto majoritário, com a aplicação de um peso igual para cada uma das predições individuais, bem como podem ser atribuídos pesos diferenciados, ou outras estratégias.

2.8 Considerações

Neste capítulo, puderam ser verificados os fundamentos teóricos que abrangem os modelos de identificação da personalidade, a partir do texto. Na sequência da conceituação da Personalidade, foram delineados os caminhos que demonstram a importância da utilização de um modelo conceituado de classificação da personalidade de um indivíduo, sendo adotado para a presente pesquisa, o modelo BIG FIVE. Por meio do estudo dos processos de identificação manual de personalidade, com a utilização de inventários de avaliação, foi identificado o questionário utilizado no experimento realizado. A partir dos conceitos da Computação da Personalidade, abrangendo o Reconhecimento Automático de Personalidade, foram verificados os meios indicados para esta identificação, com enfoque na utilização das pistas textuais para identificação dos traços de personalidade. O estudo das relações entre a Educação e a Personalidade permitiu

identificar as diversas áreas onde o processo de identificação da personalidade pode colaborar na área de Educação, reforçando a motivação para a presente pesquisa.

A análise dos modelos computacionais para a identificação da personalidade reforçou o conceito da utilização do conjunto formado por Base de Dados, Representação e Classificação, na composição do modelo apresentado. O estudo sobre as bases disponíveis para utilização em treinamento de classificadores, dentro do escopo desta pesquisa, propiciou a escolha das duas bases utilizadas no experimento realizado. O estudo dos léxicos como instrumentos utilizados para representação da personalidade presente no texto, culminou com a escolha da representação por meio do léxico LIWC como principal forma utilizada no modelo apresentado. Além disto, a utilização de duas formas complementares, baseadas em PLN, foi verificada, com a opção pelas técnicas de nGRAM e *Word2Vec*.

O conceito de Aprendizado de Máquina foi utilizado no modelo desta pesquisa, sendo de fundamental importância a escolha das técnicas para realização de comparação dos classificadores que foram utilizados. Além disto, foram apresentados os principais classificadores utilizados nos processos de classificação de texto, representado pelas técnicas já descritas, bem como as ferramentas que podem ser utilizadas. O estudo dos classificadores, bem como a investigação de trabalhos anteriores realizados com identificação de texto de forma mais genérica, culminou com a escolha da técnica de classificação *Ensemble* no modelo IP3 proposto neste projeto de pesquisa.

3 Educação e Personalidade

Com o objetivo de verificar as iniciativas existentes na literatura que abordem a relação da Personalidade com a Educação, verificando em que áreas a identificação de personalidade está sendo empregada com objetivos educacionais. Este capítulo inicia com a descrição do método de pesquisa utilizado, seguindo com uma análise crítica sobre os artigos selecionados e finalizando com as considerações sobre os resultados verificados e a contextualização da presente pesquisa em relação aos demais artigos.

Para a realização desta revisão foram utilizados os seguintes estágios: identificação dos critérios de inclusão e exclusão, bases de dados, busca de estudos relevantes, análise crítica, extração de dados e resumo . No restante desta seção são detalhados estes estágios e os resultados obtidos.

3.1 Critérios

Os critérios de inclusão de artigos, adotados nesta revisão, abrangem pesquisas que envolvam a relação de educação com a personalidade. Nos critérios de exclusão, foram descartadas publicações duplicadas, convites para eventos e pesquisas que não utilizem um modelo de classificação de personalidade. Os critérios de qualidade, definidos para esta revisão, abrangem a credibilidade, a relevância e o rigor do estudo. Como credibilidade, são verificadas se as descobertas são bem apresentadas e significativas. A relevância procura verificar a utilidade das descobertas apresentadas para a pesquisa científica na área. O rigor verifica se a abordagem utilizada é completa e adequadamente aplicada aos métodos do estudo.

3.2 Bases de Dados

A estratégia de localização de fontes primárias incluiu conceituadas bases digitais relevantes para a área de Computação e buscas manuais em periódicos e conferências nacionais. As bases digitais utilizadas foram:

- ACM Digital Library
- CAPES
- IEEE Xplore
- Science Direct (*Computer in Human Behavior*)
- SCOPUS
- Web of Science

3.3 Busca de Estudos

Com o objetivo de investigar o estado da arte na área de Educação e Personalidade, foi elaborado uma *string* de busca, para obtenção de uma quantidade significativa de artigos, para formação do acervo inicial. A busca por artigos primários, foi realizada utilizando as palavras: "*personality*" e "*education*". A Figura 3.1 ilustra a *string* adotada na revisão.

personality + **education**

Figura 3.1: *String* de Busca

A aplicação da *string* de busca na biblioteca digital "ACM Digital Library" (*Association for Computing Machinery*) obteve um retorno de 620 artigos. Este mesmo tipo de busca, aplicado na base "Science Direct" (Elsevier) obteve um retorno de 94.386 artigos. Para este caso específico, foi aplicado um refinamento de busca, adicionando na *string* os termos "*learning*" e "*identification*", o que resultou em um universo de 300 artigos. O *string* original aplicado nas base "IEEE Xplore" (*Institute of Electrical and Electronics Engineers*) propiciou o retorno de 410 artigos. Na pesquisa realizada na base "Web of Science" (Clarivate Analytics), anteriormente denominada "Web of Knowledge", foi obtido um conjunto de 8.184 artigos. Sendo aplicado um refinamento nesta busca, com a adição dos termos "*learning*" e "*identification*", a quantidade de artigos retornados foi de 41. E na base "SCOPUS" (Elsevier) foram incluídos os termos "*learning*" e "*traits*", obtendo 352 artigos. Os valores referentes a quantidade de artigos foi atualizada em 12 de março de 2018. A busca nestas bases digitais mundiais, resultou em um acervo de 1722 artigos. Evidencia-se uma curva de crescimento de publicações ao longo dos últimos anos, indicando um interesse crescente dos pesquisadores no tema escolhido para a pesquisa.

Com a aplicação de critérios de inclusão inicial, observando o título e as palavras chave, foram descartados 1488 artigos. Estes critérios de inclusão, levaram em consideração artigos que não tivessem relação direta com a abordagem da utilização da identificação da personalidade no processo educacional. Com isto o acervo obtido foi de 235 artigos. Nestes artigos resultantes foi realizada a leitura do resumo e aplicação dos critérios de exclusão, sendo descartados 215 artigos. A Figura 3.2 ilustra as etapas intermediárias do processo de obtenção da base primária. Como resultado deste processo foi obtida uma base primária composta por 20 artigos.

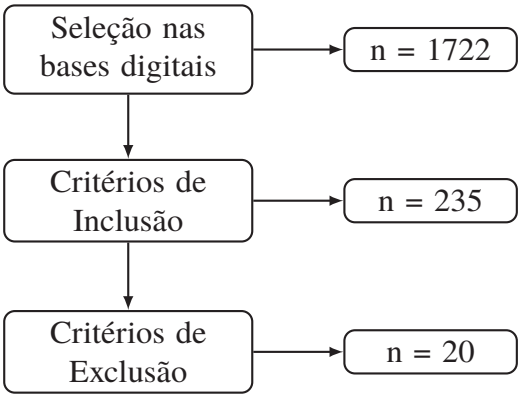


Figura 3.2: Etapas do Processo de Seleção de Artigos

3.4 Análise Crítica

Esta seção apresenta uma análise sobre os artigos selecionados, separados por tópicos, indicando seus pontos fortes e fracos, realçando os pontos principais que estão relacionados com a presente pesquisa.

3.4.1 Desempenho Acadêmico

Na pesquisa conduzida por (Blickle, 1996) em dois grupos com 139 e 92 estudantes alemães, foi verificada a relação entre os traços de personalidade, estratégias de aprendizado e performance acadêmica. A identificação do perfil de personalidade foi obtida com a aplicação de uma versão alemã do NEO-PI-R (Ostendorf e Angleitner, 1994). Os resultados indicaram que as dimensões *Conscientiousness* e *Openness* apresentaram uma forte correlação com as estratégias de aprendizado e consequente Performance Acadêmica.

O estudo empírico conduzido por Salleh et al. (2010) em universitários na Nova Zelândia, procurou verificar o efeito da dimensão *Conscientiousness* no ensino de programação em pares de alunos. Os autores concluíam que esta dimensão não apresentou grande correlação no desempenho dos pares de estudantes, mas a dimensão *Openness* demonstrou uma significativa correlação.

No estudo realizado com 133 estudantes holandeses, Kappe e van der Flier (2010) avaliaram a validade preditiva dos traços de personalidade do modelo BIG FIVE em relação à Performance Acadêmica, utilizando como ferramenta de identificação o NEO-FFI. Foram utilizadas para a correlação, cinco critérios de aprendizado: leituras em classe, treinamentos de perfil, projetos em grupo, estágio e trabalho final de graduação. Os resultados obtidos indicam que a dimensão *Conscientiousness* é um importante fator de previsão da Performance Acadêmica, nos cinco critérios adotados, e o *Neuroticism* é positivamente relacionado quando as condições de avaliação são menos estressantes.

O estudo conduzido por Di Giunta et al. (2013) verificou a relação entre os traços de personalidade, utilizando as dimensões *Conscientiousness* e *Openness*, e o desempenho acadêmico, por meio de um experimento realizado com 426 estudantes adolescentes italianos. O estudo comprovou a relação entre estes dois traços e os resultados alcançados pelos estudantes.

O estudo realizado por Weber (2015) verificou a relação entre a personalidade e o aprendizado de alunos, em um experimento realizado com 202 estudantes americanos. Foi utilizado o modelo de identificação *True Colors*, desenvolvido por Don Lowry, baseado no modelo de Myers & Briggs Myers et al. (1985) e no trabalho sobre temperamento de Keirsey e Bates (1984), que classifica a personalidade nas dimensões azul, dourada, verde e laranja. O autor concluiu que a identificação do perfil de personalidade está relacionada com o aprendizado dos alunos e esta identificação pode ajudar o professor na condução do processo educacional.

A pesquisa conduzida por Altanopoulou e Tselios (2015), em um universo de 85 estudantes universitários na Grécia, procurou verificar a relação entre as dimensões de personalidade do BIG FIVE e a Performance Acadêmica. Utilizando um ambiente baseado em *Wiki* para verificação da performance, o estudo evidenciou que a maior correlação foi verificada com a dimensão *Conscientiousness*, que o *Extraversion* demonstrou correlação negativa.

A relação entre os traços de personalidade, estilos de pensamento e performance acadêmica foi objeto do estudo realizado por Du et al. (2017). No experimento realizado com 135 alunos, no ensino da Linguagem de Programação C, foi observado que a dimensão *Agreeableness* tem um efeito positivo no ensino da linguagem de programação, mas o mesmo não ocorre com as demais dimensões.

3.4.2 Estilos de Aprendizagem

A pesquisa realizada por Donche et al. (2013) com 1.126 estudantes de ensino superior na Bélgica, teve como objetivo avaliar o impacto da personalidade e da motivação acadêmica nas estratégias de aprendizado. O perfil de personalidade no modelo BIG FIVE foi obtido por meio da aplicação do formulário NEO-FFI em holandês (Hoekstra et al., 1996), a identificação dos Estilos de Aprendizagem com a utilização do questionário ILS (Felder e Soloman, 2006) e a Motivação Acadêmica com a versão holandesa do *Academic Motivation Scale* (AMS) (Vallerand et al., 1992). Este estudo identificou as relações entre o perfil de personalidade e a estratégicas de aprendizado, com ênfase para as dimensões *Conscientiousness*, *Neuroticism* e *Openness*.

O estudo conduzido por Carro e Sanchez-Horreo (2017) verificou a influência da Personalidade e dos Estilos de Aprendizagem na Educação Colaborativa, em um estudo de caso realizado com estudantes de ensino superior de Engenharia de Computação na Espanha. Os dados dos alunos foram coletados durante a realização de atividades em ambiente virtual, com a aplicação do questionário NEO-FFI (Costa e McCrae, 1989), para a identificação do perfil de personalidade de acordo com o modelo BIG FIVE, em conjunto com a identificação de inteligência por meio do teste *Primary Mental Abilities* (PMA) (Thurstone, 1948) e dos estilos de aprendizagem com o modelo Felder-Silverman (Felder et al., 1988). Com base nestas informações, foram avaliadas, dentre outros quesitos, as correlações entre o perfil de personalidade e o desempenho, de forma individual ou em grupo, apresentando maior relevância em relação às dimensões *Agreeableness*, *Conscientiousness* e *Openness*. Este trabalho demonstra que a identificação do perfil de personalidade pode indicar os estudantes, ou grupos, com possibilidade potencial de falhar nas atividades, bem como colaborar na construção de sistemas adaptativos, utilizando a formação de grupos dinâmicos.

Na pesquisa apresentada por Cohen e Baruth (2017), é estudada a correlação entre a personalidade do estudante e a satisfação acadêmica, baseada nas evidências sobre o relacionamento entre a personalidade e o estilo de aprendizagem auto regulado (SRL), em um universo de 72 estudantes em um ambiente de aprendizagem *online*. Neste estudo foi evidenciado que as dimensões *Conscientiousness* e *Openness* evidenciam significativamente a satisfação dos estudantes e que é possível caracterizar grupos de estudantes *online*, sendo que estudantes com traços de personalidade semelhantes irão preferir canais de comunicação similares.

Um estudo sobre a utilização dos aspectos psicológicos, e da personalidade, dos alunos e suas aplicações na Informática na Educação é apresentado por Aguiar (2017). Neste estudo é apresentado o cenário das iniciativas brasileiras na utilização da personalidade na educação, concluindo que dentro da área do ensino da computação, existe uma grande variedade de contextos que podem se beneficiar dos conceitos de estilos de aprendizagem, emoções e personalidade.

3.4.3 Sistemas Adaptativos

A pesquisa conduzida por Al-Dujaily e Ryu (2008), com 33 estudantes na Nova Zelândia, procurou verificar como os diferentes perfis de personalidade, especificamente extrovertidos e introvertidos, se relacionam com o sequenciamento de atividades em um sistema de *e-learning*. O estudo comprovou que a correlação entre a performance dos grupos foi afetada pelo sequenciamento adaptado ao estilo de personalidade.

O estudo dos efeitos da identificação do perfil de personalidade no projeto de sistemas adaptativos em *e-learning* foi realizado por Kim et al. (2013), em um universo de 85 estudantes universitários na Nova Zelândia. Utilizando o modelo MBTI para identificação de personalidade e sistemas tutores inteligentes para o ensino da linguagem LISP. O estudo foi focado na verificação dos resultados decorrentes da adaptação do modelo do usuário, de acordo com o perfil introvertido

ou extrovertido e que a utilização do modelo correspondente ao perfil identificado para o estudante, acarreta melhores resultados nas atividades acadêmicas.

A pesquisa conduzida por Farias et al. (2013) aborda a apresentação de conteúdo em sistemas adaptativos, voltado ao ensino da computação, baseado no perfil de personalidade dos alunos. Os autores propõem um modelo conceitual, para ser aplicado a ambientes virtuais de aprendizagem, em que as regras de navegação são baseadas no perfil identificado do aluno.

A aplicação da identificação de personalidade na recomendação de estratégias pedagógicas personalizadas é apresentado por Melo et al. (2017). Os autores apresentam um modelo de estudante, no escopo de um sistema tutor inteligente. No experimento realizado com 172 alunos universitário em Minas Gerais, foi verificado que nos 33 alunos que foram selecionados para receber a apresentação dos resultados obtidos com a aplicação do modelo, 80% concordaram totalmente com as estratégias pedagógicas apresentadas e 20% concordaram parcialmente, sendo que nenhum aluno discordou dos resultados apresentados.

3.4.4 Outras Áreas

O trabalho de Porto et al. (2011) aborda a identificação da personalidade através do teclado como uma técnica para gerar informações para sistemas de recomendação em ambientes educacionais. Os autores realizam a proposição de uma nova técnica de recomendação, baseada na personalidade dos alunos.

Com o objetivo de observar os resultados de um ambiente EaD em relação aos traços de personalidade dos estudantes, Varela et al. (2012) realizou um experimento envolvendo 132 estudantes universitários americanos. Os autores concluem que a personalidade é uma variável digna de consideração nas configurações *online* de ensino, indicando o aprimoramento de traços mais restritos de identificação além dos traços amplos, como apresentados pelo BIG FIVE.

A análise do impacto da personalidade em sistemas de ensino baseado em MOOC *(Massive Open Online Learning*) foi verificado por Chen et al. (2016) em um estudo conduzido com 23.622 que realizaram a inscrição no curso. Deste total, cerca de 40% tiveram algum engajamento correspondente a visualização de no mínimo um vídeo. Foi oferecido no início do curso a oportunidade de preencher um formulário *online* para identificação do perfil de personalidade (BIG FIVE), sendo que 1.356 alunos completaram o formulário. Os resultados apontaram que as dimensões *Conscientiousness* e *Openness* apresentaram maiores correlações com as atividades monitoradas no experimento.

Um estudo sobre a retenção de estudantes de engenharia foi desenvolvido por Hall et al. (2015) em um grupo de 277 estudantes de uma universidade americana. O perfil de personalidade foi obtido com a aplicação do NEO-FFI. Foi verificado que a dimensão *Conscientiousness* apresentou correlação direta com a retenção dos estudantes.

3.5 Considerações

Esta revisão de literatura objetivou demonstrar a relevância da realização da identificação da personalidade em ambientes educacionais, em diversas áreas. O Desempenho Acadêmico demonstrou ser a área de maior concentração de pesquisas que utilizam o perfil de personalidade, como balizador de estratégias de ensino. Também foram identificadas iniciativas que relacionam a personalidade com: Estilo de Aprendizagem, Sistemas Adaptativos, Sistemas de Recomendação e na área de Retenção.

Foi verificado que a maior parte das iniciativas obtém o perfil de personalidade dos alunos de forma manual, ou ainda, com a utilização de formulários *online*, demonstrando um

espaço a ser coberto pela identificação automática e não intrusiva. Acredita-se que a presente pesquisa contribui no campo da Educação por oferecer esta alternativa de identificação automática da personalidade dos alunos, que pode ser utilizada como um instrumento que ofereça subsídios sobre o perfil dos alunos para utilização pelos envolvidos nos processo de ensino-aprendizagem, com a utilização de interfaces adaptativas e outras técnicas relacionas, que considerem as diferenças dos alunos.

Como resultado parcial da presente pesquisa, ocorreu a publicação do artigo "Identificação de estilo de aprendizagem: Um modelo de inferência automatizado baseado no perfil de personalidade identificado nos textos produzidos pelo aluno" (Buiar et al., 2017), no XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017). Este trabalho apresenta o modelo de identificação automática de personalidade de alunos, aplicado na realização da inferência do Estilo de Aprendizagem dos alunos, de acordo com o modelo de Felder e Silverman (Felder et al., 1988). A partir do desenvolvimento do modelo IP3, a identificação automática do perfil de personalidade dos alunos pode ser aplicado em diversas áreas da Educação, como também demonstrado no trabalho *Detecção automática de traços de personalidade e recomendação de agrupamento com o modelo Big Five*. Este trabalho foi aceito para publicação no XXIX Simpósio Brasileiro de Informática na Educação (SBIE 2018), realizado em conjunto com a pesquisadora Taís Ferreira (UFU), apresentando a utilização desta identificação, como instrumento de apoio para a formação de grupos de colaboração, em sala de aula.

4 Identificação de Perfil de Personalidade Baseado em Texto

Este capítulo tem como objetivo apresentar uma revisão da literatura relacionada com a Identificação do Perfil de Personalidade Baseado em Texto. A partir dos resultados obtidos com os artigos selecionados, são apresentados os modelos de personalidade mais utilizados, as bases de dados adotadas, os idiomas utilizados, as formas de representação do texto e as técnicas de classificação, adotados nos experimentos investigados. Na sequência é realizada uma apresentação das acurácias descritas nos artigos selecionados, relacionadas com os métodos adotados pelos pesquisadores. Após uma análise crítica dos estudos selecionados, são apresentadas as considerações sobre os resultados obtidos e o posicionamento da presente pesquisa em relação às demais.

4.1 Critérios

Os critérios de inclusão de artigos, adotados nesta revisão, abrangem as publicações especificamente relacionadas à identificação automática de personalidade que atendam aos critérios de qualidade definidos para a revisão. Nenhuma restrição sobre a data de publicação foi aplicada. Como critérios adicionais de inclusão, foram considerados somente os artigos que possam estar relacionados à utilização de um modelo computacional para classificação automática de textos.

Nos critérios de exclusão, foram descartadas as publicações que não apresentaram resultados empíricos, artigos relacionados com outras pesquisas já selecionadas e que não apresentavam um modelo computacional. Os critérios de qualidade, definidos para esta revisão, abrangem a credibilidade, a relevância e o rigor do estudo. Como credibilidade, são verificadas se as descobertas são bem apresentadas e significativas. A relevância procura verificar a utilidade das descobertas apresentadas para a pesquisa científica na área. O rigor verifica se a abordagem utilizada é completa e adequadamente aplicada aos métodos do estudo.

4.2 Questões de Pesquisa

Considerando que as questões de pesquisa devem exemplificar os objetivos do mapeamento realizado, foram elaboradas as seguintes questões, a serem verificadas na investigação da literatura:

QID_1 - Quais os modelos de traços de personalidade são considerados ?

QID_2 - Quais as bases de dados são utilizadas para validação dos modelos ?

QID_3 - Em quais idiomas de texto os experimentos foram realizados ?

QID_4 - Quais as formas de representação do texto foram utilizadas ?

QID_5 - Quais as técnicas de classificação que foram adotadas ?

4.3 Bases de Dados

A estratégia de localização de fontes primárias incluiu conceituadas bases digitais relevantes para a área de Computação e buscas manuais em periódicos e conferências nacionais. As bases digitais utilizadas foram:

- ACM Digital Library
- CAPES
- IEEE Xplore
- Science Direct (*Computer in Human Behavior*)
- SCOPUS

4.4 Busca de Estudos

Tendo como referência, a pergunta da presente pesquisa, foi elaborado uma *string* de busca, para obtenção de uma quantidade significativa de artigos, para formação do acervo inicial. Foram descartadas publicações não relacionadas a processos ou modelos computacionais, visto o grande número de artigos relacionados com identificação de personalidade, voltados somente para a área de Psicologia.

A busca por artigos primários, foi realizada utilizando as palavras: "*personality*", relacionada com o foco principal da pesquisa; "*classification*", que indica um delimitador de escopo para os modelos procurados e a palavra "*text*", para delimitar o sub conjunto relacionado ao tipo de classificação desejado na busca, objetivando eliminar outras modalidades, como identificação por fala, por exemplo. A Figura 4.1 ilustra a *string* adotada na revisão.

personality + **classification** + **text**

Figura 4.1: *String* de Busca

A aplicação da *string* de busca na biblioteca digital "ACM Digital Library" (*Association for Computing Machinery*) obteve um retorno de 146 artigos. Este mesmo tipo de busca, aplicado na base "Science Direct" (Elsevier) obteve um retorno de 9.755 artigos. Para este caso específico, foi aplicado um refinamento de busca para incluir somente o periódico "*Computer in Human Behavior*", o que resultou em um universo de 124 artigos. O *string* original aplicado na base "IEEE Xplore" (*Institute of Electrical and Electronics Engineers*) propiciou o retorno de 29 artigos. Na pesquisa realizada na base "Web of Science" (Clarivate Analytics), anteriormente denominada "Web of Knowledge", foi obtido um conjunto de 87 artigos. E na base "SCOPUS" (Elsevier) a coletânea obtida foi de 144 artigos primários. Os valores referentes à quantidade de artigos foi atualizada em 12 de março de 2018.

A busca nestas bases digitais mundiais, resultou em um acervo de 530 artigos, aos quais foram acrescentados mais 17 artigos oriundos de buscas específicas em periódicos e congressos

nacionais, formando um acervo candidato de 547 artigos. Evidencia-se uma curva de crescimento de publicações ao longo dos últimos anos, indicando um interesse crescente dos pesquisadores no tema escolhido para a pesquisa.

Deste total de artigos primários, 78 estavam duplicados e 18 eram convites para participação em eventos e abertura de congressos, resultando numa redução do escopo para 453 artigos candidatos. Com a aplicação de critérios de inclusão inicial, observando o título e as palavras chave, foram descartados 351 artigos. Estes critérios de inclusão, levaram em consideração artigos que possam estar relacionados a um modelo computacional para classificação automática de textos, considerando o perfil de personalidade, bem como as pesquisas que apresentaram resultados empíricos. Sendo assim, a base primária candidata ficou restrita a 100 artigos.

Os artigos candidatos resultantes, foram novamente filtrados, utilizando como critério de exclusão: artigos duplicados oriundos de uma mesma pesquisa; pesquisas que não envolveram a utilização de um modelo computacional; modelos de classificação não relacionados ao perfil de personalidade; artigos que não apresentaram resultados de experimentos. Considerando estes filtros, foram excluídos 75 artigos da base candidata. A Figura 4.2 ilustra as etapas intermediárias do processo de obtenção da base primária. Como resultado deste processo foi obtida uma base primária composta por 25 artigos.

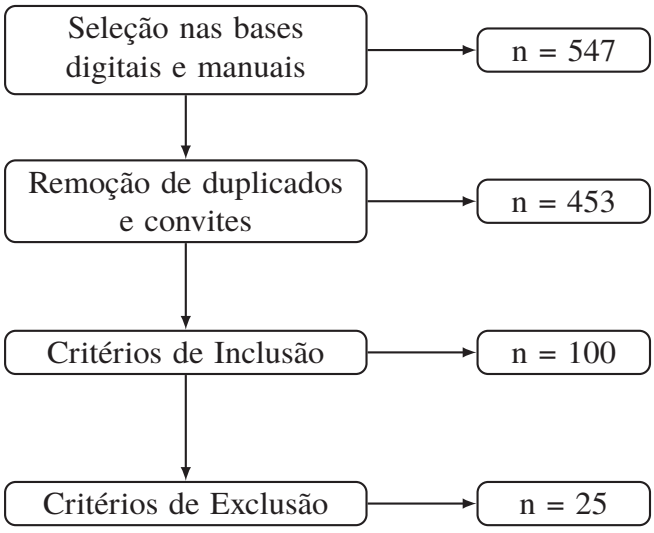


Figura 4.2: Etapas do Processo de Seleção de Artigos

4.5 Resultados

Esta seção descreve os resultados obtidos com a análise dos artigos que formam o acervo primário. Com base na leitura destes artigos foi realizada uma referência cruzada entre as questões da pesquisa e o conteúdo dos artigos selecionados. Para cada uma das questões de pesquisa, QID_1 a QID_5, foi verificada a ocorrência dos itens indicados em cada questão, sendo apresentado o resultado estatístico e as considerações sobre os valores encontrados.

4.5.1 Modelos de Traços de Personalidade

A verificação do modelo de identificação automática de traços de personalidade utilizadas nos experimentos, oferece um direcionamento sobre qual modelo seria mais adequado a ser investigado na pesquisa. Com base na questão

"*QID_1 - Quais os modelos de traços de personalidade são considerados ?*"

foi realizada a investigação dos modelos utilizados nas pesquisas. A Figura 4.3 apresenta a quantidade de artigos que utilizam cada um dos modelos encontrados. É notória a predominância do modelo BIG FIVE em 20 dos artigos selecionados, o que corresponde a 80% da amostra. As pesquisas de Luyckx e Daelemans (2008), Komisin e Guinn (2012) e Ezpeleta et al. (2016) optaram por utilizar o modelo de personalidade MBTI (Myers et al., 1985), ao passo que Saez et al. (2014) utilizou o modelo de Três Fatores de Eysenck (Eysenck, 1947). O trabalho de Minamikawa e Yokoyama (2011) utilizou o modelo *Ecogram*, originalmente proposto por Dusay (1972). Não foram identificados argumentos sobre os critérios utilizados pelos pesquisadores para seleção de um modelo ou outro, de personalidade nos estudos, mas sim, considerações sobre a necessidade de utilização de um modelo como referência, havendo um foco maior no modelo de identificação a ser utilizado, ou seja, como encontrar os traços de personalidade registrados nos textos utilizados nos experimentos.

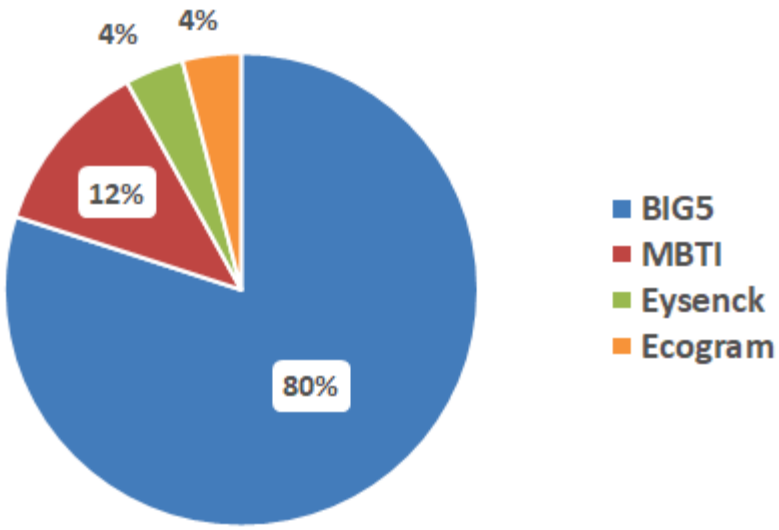


Figura 4.3: Modelos de Personalidade

4.5.2 Bases de Dados

A investigação das bases utilizadas pelos pesquisadores oferece uma importante referência para a realização de experimentos que possam ser comparados com as diferentes propostas encontradas na literatura. Com base na questão

"*QID_2 - Quais as bases de dados são utilizadas para validação dos modelos ?*"

foi realizada a investigação das bases de dados utilizadas. A Figura 4.4 apresenta a quantidade de artigos que utilizaram cada tipo de base em seus experimentos. Nos casos em que mais de

uma base foi utilizada, o artigo é incluído mais de uma vez. A referência à base *Facebook* está associada a bases obtidas a partir da plataforma em questão, não incluindo a base *myPersonality*, para a qual foi utilizado um item próprio.

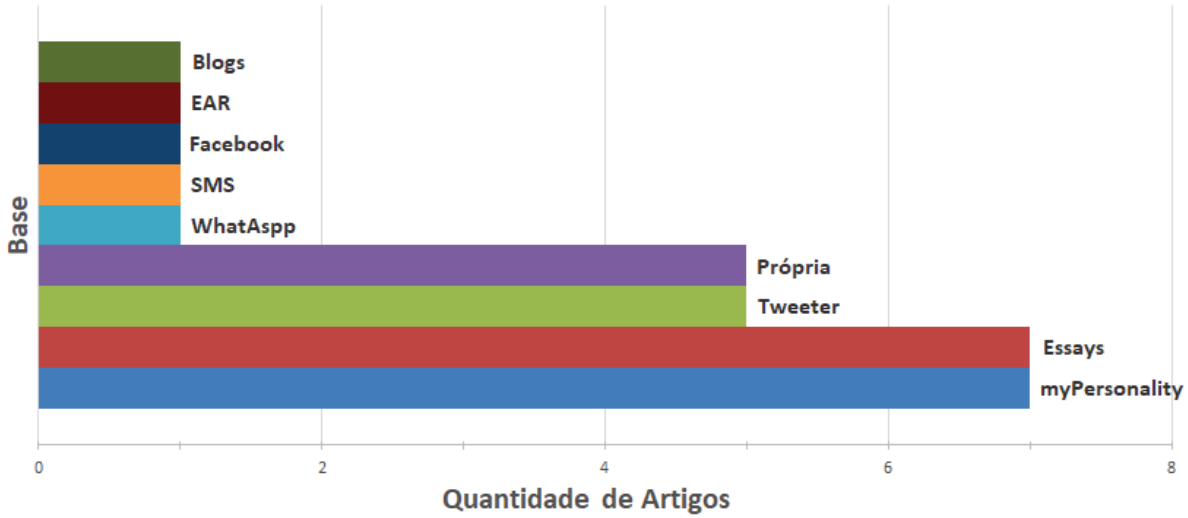


Figura 4.4: Tipos de Bases

Pode ser observada a predominância das bases ESSAYS e *myPersonality*, dentre as mais utilizadas pelos experimentos dos artigos selecionados. Como bases próprias são considerados os casos em que os pesquisadores realizaram o experimento utilizando somente bases obtidas durante a pesquisa. Nestes casos, de bases próprias, foram verificadas bases obtidas em ambientes educacionais, como nas pesquisas de Luyckx e Daelemans (2008), Komisin e Guinn (2012) e Ferreira e Fernandes (2017), nas quais os alunos realizaram a elaboração de um texto solicitado pelos pesquisadores. A pesquisa de Wei et al. (2017) também foi realizada em ambiente educacional e utilizou uma base própria, sendo que os textos utilizados foram obtidos das publicações que os alunos realizaram no microblog chinês *Sina Weibo*. Em nenhuma dos artigos selecionados foi verificada a utilização de uma base de textos oriundas de atividades acadêmicas do dia a dia dos alunos. Também foi observada a obtenção por diversos pesquisadores de textos oriundos de redes sociais diversas, sendo que em muitos casos, além do texto em si, são utilizados outros indicadores para a formação da base para classificação.

4.5.3 Idiomas do Texto

A verificação do idioma do texto utilizado pode identificar a utilização ou não de idiomas diversos nas pesquisas de identificação da personalidade. Como quesito idioma, pretende-se investigar o idioma da base de dados que foi utilizada no experimento e não o idioma utilizado na publicação da pesquisa. Na busca dos resultados oriundos da questão

"*QID_3 - Em quais idiomas de texto os experimentos foram realizados ?*"

foi feita a verificação do idioma dos textos usados pelos pesquisadores, na realização dos experimentos de identificação de perfil de personalidade, sendo que os resultados estão apresentados na Figura 4.5.

Fica evidente a predominância do idioma inglês, presente em 76% dos experimentos identificados. Foi também observada a presença de dois artigos nacionais, Nunes et al. (2013) e Ferreira e Fernandes (2017), que realizaram os experimentos em português e quatro experimentos

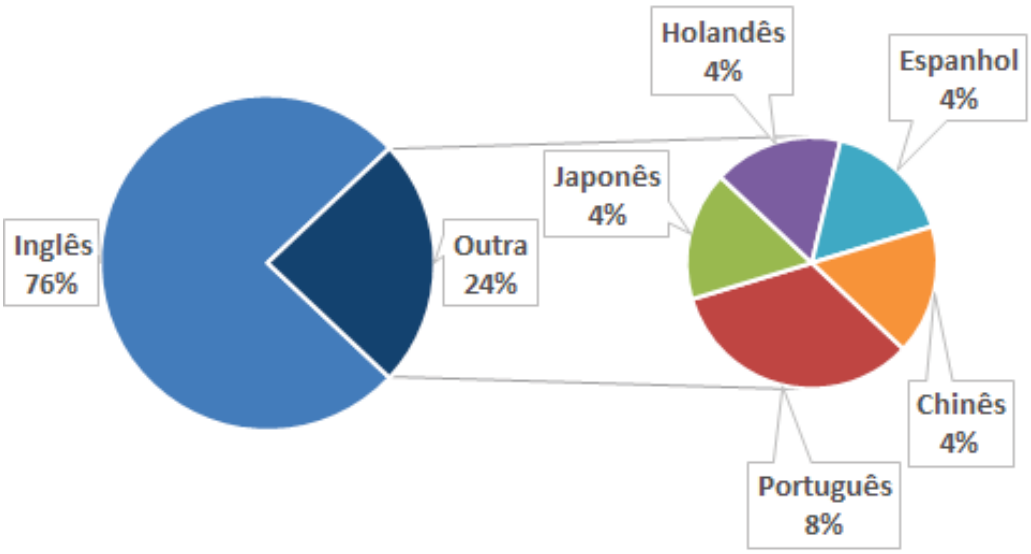


Figura 4.5: Idioma do Texto utilizado no Experimento

isolados realizados em idiomas locais: holandês (Luyckx e Daelemans, 2008); japonês (Komisin e Guinn, 2012); espanhol (Saez et al., 2014) e chinês (Wei et al., 2017).

4.5.4 Formas de Representação

A forma na qual os pesquisadores realizaram a estruturação do texto com o objetivo de aplicação de técnicas de classificação, cujos resultados podem ser visualizados na Figura 4.6, foi verificado com o questionamento

"*QID_4 -Quais as formas de representação do texto foram utilizadas ?*"

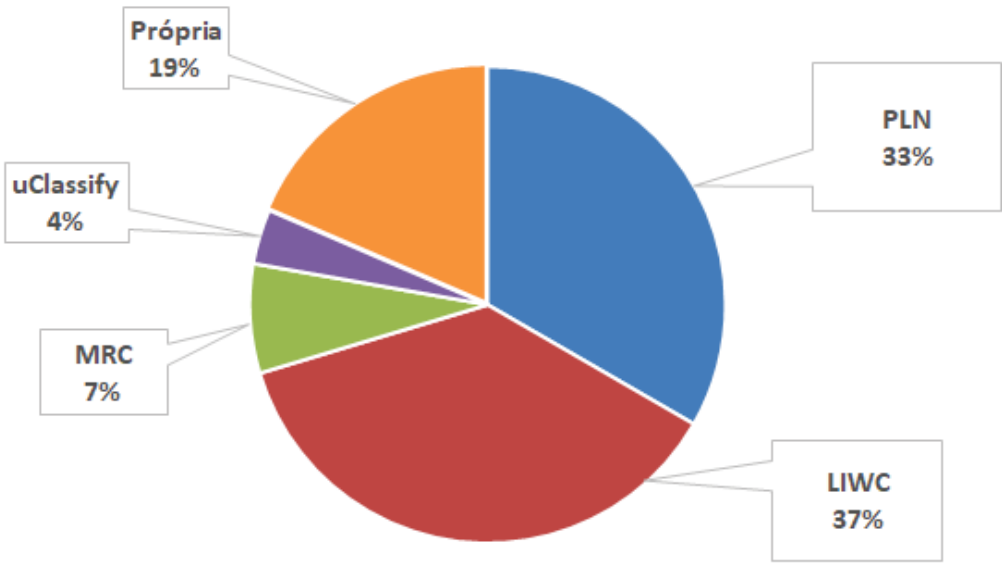


Figura 4.6: Formas de Representação do Texto

A quantidade de artigos que utiliza cada um destes métodos é apresentada, sendo que nos casos do experimento em questão que utilizar mais de um método, o artigo é contabilizado mais de uma vez. Estes foram os casos das pesquisas de: Mairesse et al. (2007) que utilizou o LIWC e o MRC; Komisin e Guinn (2012) com a utilização de LIWC e PLN e Poria et al. (2013) na utilização de LIWC e MRC.

Foi verificada a grande predominância da utilização do LIWC (37%) e de métodos de PLN (30%), sendo que esta categoria inclui as suas diversas variantes como nGRAM e *Word2Vec*. Na categoria "Própria" estão agrupadas as técnicas de extração específicas desenvolvidas pelos autores, nos casos de: Lima e Castro (2014), Saez et al. (2014); Kalghatgi et al. (2015) e Liu et al. (2016). Foi verificada a utilização da ferramenta *uClassify* (Kågström et al., 2013) na pesquisa de Ezpeleta et al. (2016).

4.5.5 Técnicas de Classificação

Outro ponto importante a ser verificado na literatura recente está relacionado às técnicas de classificação utilizadas nos processos de aprendizado de máquina com a finalidade de identificação do perfil de personalidade, a partir dos traços identificados no texto. Como resposta a questão

"*QID_5 - Quais as técnicas de classificação que foram adotadas ?*"

foi observada uma grande gama de tipos de classificadores e métodos associados. A Figura 4.7 apresenta a quantidade de artigos que citaram a utilização das principais famílias de classificadores. Ocorreu com bastante frequência o caso de artigos que utilizaram mais de um tipo de classificador nos experimentos realizados.

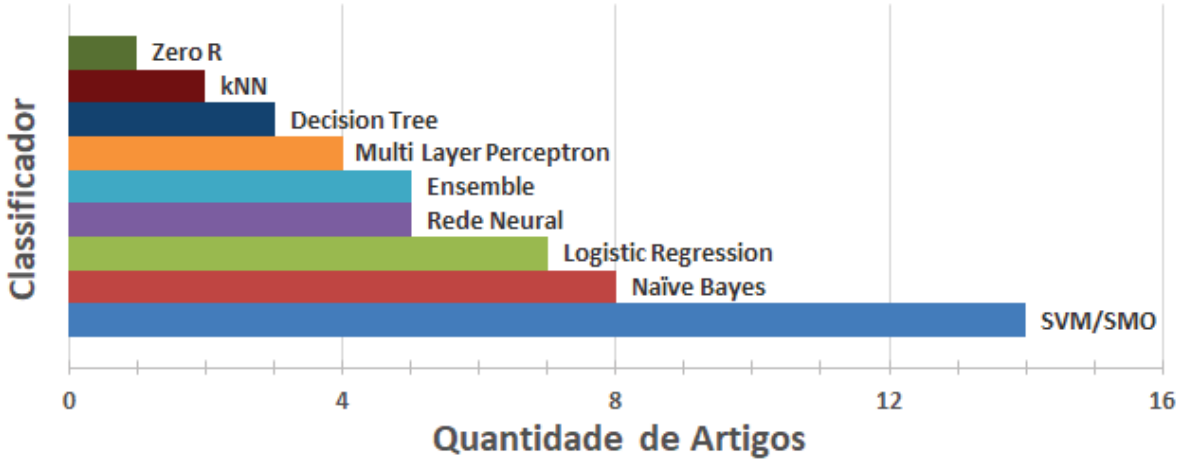


Figura 4.7: Tipos de Classificadores

Os classificadores foram agrupados de acordo com as seguintes famílias:

- kNN - *k-nearest neighbors*;
- NB - *Naïve Bayes* em seus diversos tipos;
- *Zero R*;

- SVM - *Support Vector Machine*;
- *Decision Tree*;
- *Logistic Regression*;
- MLP - *Multi Layer Perceptron*;
- *Ensemble* - abrangendo *Random Forest*, *Adaboost*, *Gradiente Boosting* e similares;
- Rede Neural - que inclui *Convolutional Neural Network* (CNN), *Recurrent Neural Network* (RNN) e similares.

Um levantamento dos resultados obtidos com os diversos tipos de classificadores foi realizado e está apresentado na Tabela 4.1. Os valores apresentados indicam os melhores resultados obtidos em cada uma das dimensões OCEAN, bem como o classificador utilizado em cada caso. Somente estão incluídos os resultados dos artigos que utilizaram as bases ESSAYS ou *myPersonality* e que indicaram a acurácia nas publicações.

Tabela 4.1: Acurácia Obtida nos Artigos Investigados

ARTIGO	BASE	REPR	OPN	CON	EXT	AGR	NEU
Mairesse et al. (2007)	ESS	LIWC	63%^{SVM}	56%^{SVM}	56%^{ADA}	56%^{SVM}	58%^{SVM}
Iacobelli e Culotta (2013)	ESS	PLN	62%^{GNB}	55%^{GNB}	56%^{GNB}	53%^{GNB}	56%^{GNB}
Tighe et al. (2016)	ESS	LIWC	61%^{SVM}	55%^{LR}	54%^{SVM}	57%^{LR}	57%^{LR}
Majumder et al. (2017)	ESS	LIWC	63%^{NN}	57%^{NN}	59%^{NN}	57%^{NN}	59%^{NN}
Alam et al. (2013)	MyP	PLN	69%^{MNB}	59%^{MNB}	59%^{MNB}	58%^{MNB}	63%^{MNB}
Iacobelli e Culotta (2013)	MyP	PLN	54%^{GNB}	56%^{GNB}	61%^{GNB}	53%^{SVM}	48%^{GNB}
Tandera et al. (2017)	MyP	LIWC	74%^{NN}	56%^{NN}	65%^{NN}	59%^{NN}	65%^{NN}
Yu e Markov (2017)	MyP	PLN	71%^{NN}	51%^{NN}	61%^{NN}	54%^{NN}	61%^{NN}

BASE= ESSAYS^{ESS}, *myPersonality*^{MyP}. REPR = Léxico LIWC^{LIWC}, PLN^{PLN}. Classificadores= *AdaBoost*^{ADA}, *Gaussian Naïve Bayes*^{GNB}, *Logistic Regression*^{LR}, *Multinomial Naïve Bayes*^{MNB}, *Neural Network*^{NN}, *Support Vector Machine*^{SVM}.

Pode ser verificado que os melhores resultados são obtidos com a base *myPersonality*, mas deve-se levar em consideração que muitas das pesquisas que envolvem esta base, também utilizam informações de redes sociais que estão presentes nesta base, e não somente o texto. Em geral os valores de acurácia estão em torno de 60%, não apresentando variações significativas, mesmo com os diversos modelos adotados. A dimensão que demonstra uma maior acurácia nos estudos identificados é a *Openness*, condição esta que também foi verificada no experimento realizado durante a presente pesquisa.

4.6 Análise Crítica

Esta seção apresenta uma análise sobre os artigos selecionados, realçando os pontos principais que estão relacionados com a presente pesquisa. A Figura 4.8 apresenta um diagrama destes artigos, sendo que estes estão agrupados de acordo com a base utilizada e o modelo de representação do texto.

As bases utilizadas estão representadas pelos agrupamentos indicados com linha cheia, onde os experimentos que utilizam mais de uma base estão indicados de forma sobreposta, como no caso de Iacobelli e Culotta (2013), que utilizou as bases ESSAYS e *myPersonality*. A forma

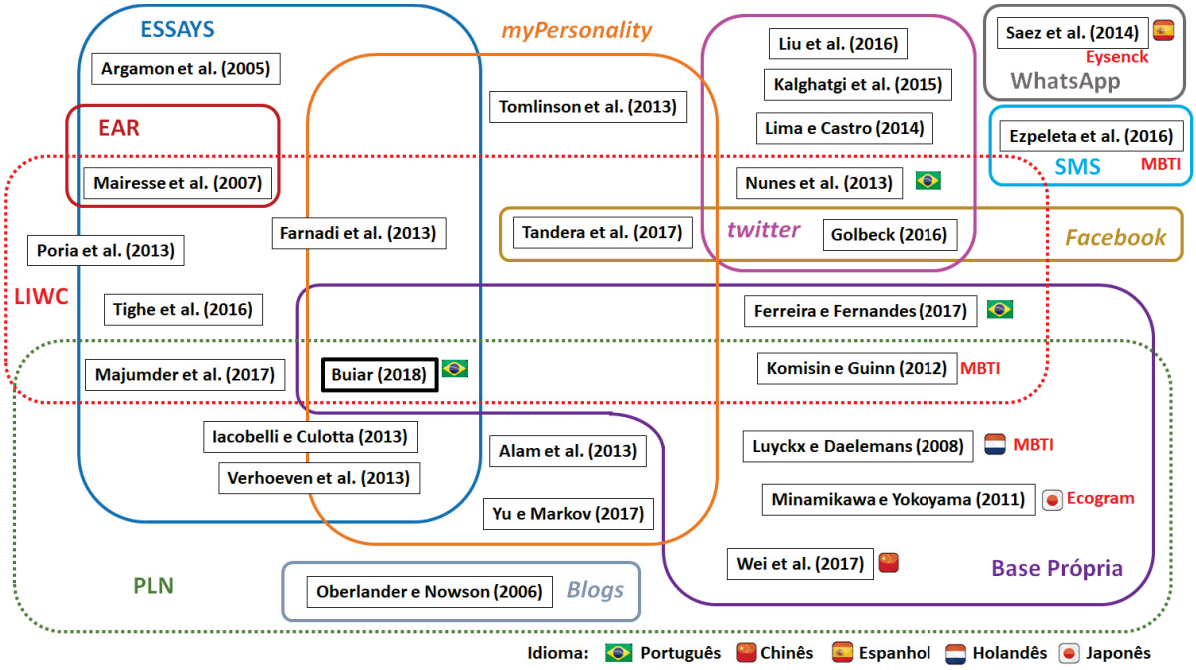


Figura 4.8: Artigos Pesquisados

de representação do texto está indicada em linha pontilhada, quando foi utilizado LIWC ou PLN, e novamente, nos casos em que ambas as formas foram utilizadas, ocorre a sobreposição, como no caso de Komisin e Guinn (2012). As demais formas de representação utilizadas, não estão destacadas neste diagrama. O modelo de personalidade, para os casos em que não foi utilizado o BIG FIVE, está indicado em vermelho, abaixo da referência da pesquisa, como no caso de Saez et al. (2014), que utilizou o modelo Eysenck. A maioria das pesquisas utilizou o experimento com textos em inglês, sendo que as exceções estão indicadas conforme a legenda no rodapé da figura, como no caso de Nunes et al. (2013) que utilizou o idioma português.

No experimento da presente pesquisa, indicada em destaque no diagrama como "Buiar (2018)", foram utilizadas as bases ESSAYS e *myPersonality* combinadas com uma base própria em português obtida a partir de textos de atividades educacionais, as formas de representação LIWC e PLN combinadas com um conjunto de classificadores de forma parametrizável, representando uma iniciativa a nosso ver inédita, na área de identificação de personalidade a partir de textos educacionais.

Na pesquisa de Argamon et al. (2005), somente foram abrangidas as dimensões *Extraversion* e *Neuroticism*. Este trabalho foi um dos precursores no estudo da relação entre os fatores léxicos e psicológicos do texto escrito e a personalidade, utilizando métodos computacionais. Utilizando uma base própria, agregando dados sociais, Oberlander e Nowson (2006) utilizaram as cinco dimensões do BIG FIVE bem como as formas de representação *unigram* e *bigram*. Os resultados apresentados, indicaram altos valores de acurácia, mas devem ser verificados com cautela, visto ter sido utilizada, uma base reduzida e específica, o que poderia induzir a resultados muito superiores, em relação aos obtidos em outros experimentos investigados na literatura. Outro ponto observado, foi que, para a obtenção de maiores valores de acurácia, os autores descartaram parte da base de dados originais, em processos sucessivos. Esta abordagem poderia levar a interpretação que os resultados obtidos, refletem somente uma parte mais apropriada da base de dados.

Outro trabalho precursor nesta área , realizado por Mairesse et al. (2007), utiliza além da base ESSAYS, uma segunda base, contendo conversas em áudio obtidas utilizando gravadores

de voz ativados eletronicamente (EAR), formando um *corpus* de cerca de 100.000 palavras de 96 participantes. No caso desta segunda base, o perfil de personalidade dos autores foi obtido com a utilização de questionário de autoavaliação, bem como avaliações feitas por 18 observadores independentes. As características do texto foram extraídas utilizando o LIWC, além de 14 características adicionais do MRC. Uma importante conclusão apresentada pelo estudo, é a demonstração de que a utilização das características extraídas com o LIWC apresenta resultados superiores, em comparação, com a utilização do MRC, ou utilizando a combinação de ambos.

Dentre os trabalhos que utilizam a base ESSAYS com o LIWC, Poria et al. (2013) utiliza também o MRC para extração de características, além de outros recursos léxicos afetivos como o *SentiNet*, *ConceptNet*, *EmoSentiNet* e *EmoSenticSpace*. Os resultados demonstraram que a combinação dos diversos léxicos afetivos apresentaram acurácia superior ao trabalho de Mairesse et al. (2007). Já o trabalho desenvolvido por Tighe et al. (2016) está voltado para a utilização das técnicas de redução de características *Information Gain* e Principal *Component Analysis* aplicadas na identificação de personalidade baseado em texto. Os resultados demonstraram que significativas reduções de dimensionalidade foram obtidas com a técnica *Information Gain*. Apesar disto, foi observado que os ganhos reais obtidos em termos de acurácia não chegam a 2%. O trabalho de Majumder et al. (2017) aplica Redes Neurais Convolucionais (CNN) para a extração das características do texto. O modelo apresentado realiza a extração de *unigram*, *bigram* e *trigram* combinados com as 84 características obtidas com o LIWC, formando um vetor de características de 684 dimensões. Este trabalho demonstrou que a utilização isolada das técnicas de nGRAM ou CNN, não apresentam resultados melhores do que a aplicação isolada do LIWC.

A pesquisa de Farnadi et al. (2013) utilizou a base *myPersonality*, mas além das informações obtidas do texto, também utilizou outras informações disponíveis nesta base, como as características das redes sociais, informações sobre frequências de inserções de mensagens ao longo do tempo e outras obtidas na base. As características do texto foram obtidas com o léxico LIWC. Uma inovação desta proposta, foi o treinamento dos classificadores utilizando outra base mais consistente, no caso a base ESSAYS. Os resultados demonstraram que é possível a utilização combinada destas bases, sendo que os valores obtidos, não apresentam grandes variações, em relação a utilização somente da base *myPersonality*.

Os trabalhos de Iacobelli e Culotta (2013) e Verhoeven et al. (2013), utilizaram as bases ESSAYS e *myPersonality* para a realização dos experimentos, sendo que a representação do texto foi realizada com PLN. A iniciativa de Iacobelli e Culotta (2013), foi o único trabalho, dentre os selecionados, que não considerou que as dimensões BIG FIVE são independentes, mas isto não acarretou significativas mudanças nos resultados, como também foi a única pesquisa que agrupou os textos de mesmos autores na base *myPersonality*, obtendo 250 documentos ao invés dos 9917 originais. Esta estratégia de agrupamento foi realizada nos experimentos iniciais da presente pesquisa, mas os resultados apresentados não demonstraram diferenças significativas em relação aos resultados obtidos com a base *myPersonality* original. Já Verhoeven et al. (2013) utilizou as bases ESSAYS e *myPersonality* de forma agregada, baseado na pesquisa de Farnadi et al. (2013), aumentando o tamanho da base de treinamento. Nos experimentos realizados na presente pesquisa, o agrupamento destas bases, para identificação do perfil de personalidade de textos de alunos, apresentou uma acurácia menor, do que a utilização somente da base ESSAYS.

Os trabalhos de Alam et al. (2013) e Yu e Markov (2017), utilizam a base *myPersonality* e representação PLN. No caso do primeiro trabalho, os autores indicam que, na época da publicação dos resultados, não havia outras referências utilizando a base *myPersonality*, sendo que atualmente já existem diversos trabalhos publicados com resultados divulgados para esta base. O trabalho de Yu e Markov (2017) apresenta a identificação da personalidade por meio de

utilização de redes neurais *fully-connected* (FC), convolucionais (CNN) e recorrentes (RNN), aplicadas sobre a base *myPersonality*. Um ponto observado neste trabalho, foi que, além da utilização do texto como fonte de informação para identificação de personalidade, as informações da rede social, disponibilizadas na base *myPersonality*, foram utilizadas pelos autores, o que representa um conjunto adicional de características. Os resultados apresentados, não refletem somente a acurácia obtida na identificação, utilizando exclusivamente texto, como o apresentado por diversos autores que também realizaram experimentos com a base *myPersonality*. Na presente pesquisa somente estão consideradas as informações obtidas nos textos da base *myPersonality*, quando os experimentos utilizam esta base, visto que objeto a ser investigado está associado a identificação da personalidade dos alunos, baseada somente em informações do texto presente nas atividades educacionais.

O trabalho de Tandera et al. (2017) utiliza técnicas de *deep learning* para identificação de personalidade no modelo BIG FIVE. São utilizadas duas bases no experimento, a *myPersonality* e uma outra base obtida pelos autores com 150 usuários do *Facebook*. Para a extração de características foi utilizado o LIWC 2015, bem como o SPLICE (Moffitt et al., 2012) e as informações da rede social presentes na base *myPersonality*. Os resultados apresentados pelo experimento demonstraram que a utilização das características adicionais de rede social e do SPLICE não apresentaram melhores resultados do que a utilização somente do LIWC. A utilização de processos de *deep learning* apresentou significativos ganhos em termos de acurácia em relação aos processos tradicionais de classificação, de acordo com os resultados demonstrados no trabalho, o que caracteriza a relevante contribuição apresentada pelos autores. Outra proposta que utiliza a base *myPersonality*, foi a conduzida por Tomlinson et al. (2013), focada na dimensão *Conscientiousness*. A obtenção de características é realizada a partir da análise semântica das informações textuais da base. O trabalho apresenta uma técnica alternativa de obtenção de características, para identificação de personalidade, a partir de texto, mas somente trata de uma das dimensões do BIG FIVE.

Além das bases ESSAYS e *myPersonality*, utilizadas pelas principais pesquisas que foram baseadas no modelo BIG FIVE, foram identificadas iniciativas que utilizaram base própria e representação com LIWC, como nos casos das iniciativas de Nunes et al. (2013) e Ferreira e Fernandes (2017), que representam uma importante contribuição por tratar de identificação de textos em português. A primeira não utilizou classificadores, mas somente apresentou a correlação entre a identificação e os dados obtidos com questionários. A segunda, procurou identificar a personalidade dos alunos para formação de grupos de colaboração. A utilização de bases consolidadas de referência e indicadores de resultados similares aos demais trabalhos encontrados na literatura, como o realizado pela presente pesquisa, poderiam facilitar o comparativo destes trabalhos com os demais.

Os trabalhos de Luyckx e Daelemans (2008), Komisin e Guinn (2012) e Ezpeleta et al. (2016) optaram pelo modelo MBTI de personalidade, que possui quatro dimensões. O primeiro foi uma iniciativa em idioma holandês, e os demais, em inglês. O segundo realizou um comparativo entre a utilização do LIWC e PLN, constatando que a utilização do LIWC apresenta resultados superiores. A utilização conjunta dos dois métodos, como foi desenvolvido no modelo apresentado pela presente pesquisa, poderia agregar mais resultados ao trabalho apresentado. A terceira apresentou a proposta de identificação de personalidade para identificação de mensagens de *spam* em textos de mensagens curtas (SMS), demonstrando uma aplicação inovadora para a identificação de personalidade a partir de textos.

O estudo de Minamikawa e Yokoyama (2011) utilizou uma base própria, selecionada de 551 autores, obtidos em *blogs* japoneses, sendo que os autores realizaram a identificação manual com um formulário denominado TEG2 (*Todai-shiki Egogram version 2*), específico para a língua

japonesa. Ao invés do tradicional BIG FIVE, os autores utilizaram o modelo *Egogram* (Dusay, 1972), focado na parte relacionada ao ego, na personalidade. As características foram obtidas com PLN, obtendo vetores de características de tamanho 50, 100, 150, 200 e 250. Também foram realizados experimentos com classificação binária, utilizando 3 classes e também 5 classes. Os melhores resultados foram obtidos na classificação binária, com vetores de características de tamanho superior a 150. Este estudo, representa mais uma interessante contribuição de identificação de personalidade em língua não inglesa. Por outro lado, fica difícil uma comparação da eficácia do método adotado, por não ser utilizada uma base mais padronizada.

A pesquisa conduzida por Wei et al. (2017) apresenta um modelo de identificação de personalidade a partir de informações heterogêneas, envolvendo o texto de *tweets*, o avatar do usuário, os *emoticons* registrados nas mensagens e as iterações entre os usuários. O grupo de usuários utilizado foi de 3.162 alunos de uma escola de medicina da China. Na parte de identificação a partir do texto foi utilizada a separação em *bag-of-words* das palavras em inglês e chinês, bem como pontuação em inglês e chinês. Além disto foi utilizada uma rede CNN para extração das características. A classificação combinada deste grande conjunto de classes de características realiza a identificação da personalidade utilizando o classificador LR. Este modelo heterogêneo apresentou promissores resultados na base utilizada, representando uma importante contribuição, mas novamente fica inviável a comparação com os outros estudos que utilizam bases consolidadas e textos somente em inglês.

O trabalho apresentado por Golbeck (2016) realiza uma análise sobre a identificação de personalidade utilizando uma ferramenta denominada Receptiv*iti* API, oferecida pela organização Receptiv*iti* fundada em 2015, sendo que um de seus co-fundadores, James Pennebaker, foi responsável pelo desenvolvimento do LIWC (Seção 2.6.3). Esta API permite a análise de um texto por meio de um acesso JSON e utiliza como referência o modelo BIG FIVE. Os pesquisadores utilizaram quatro bases de dados para realização dos ensaios, duas bases oriundas de informações obtidas das ferramentas *Twitter* e *Facebook*, respectivamente, e as outras duas obtidas junto ao projeto *myPersonality*, sendo a base com 10.000 registros e outra com 22 milhões de registros. Apesar da grande quantidade de registros obtidos na pesquisa, esta foi limitada à apresentação de dados estatísticos sobre a identificação realizada pela ferramenta, não tendo sido realizado ensaios com classificadores próprios.

As pesquisas que utilizaram bases obtidas do *twitter* incluem Lima e Castro (2014), Kalghatgi et al. (2015) e Liu et al. (2016). A primeira apresenta o modelo PERSOMA, com a proposta de ser um sistema de predição de personalidade baseado na análise de dados obtidos em mídias sociais, sendo utilizadas três bases públicas de *tweets*, *Obama–McCain Debate* (OMD), *Sanders* e SemEval2013, previamente rotuladas com informações de sentimentos. Para a representação de texto, foram utilizados dados estatísticos obtidos no texto e informações de redes sociais. Os melhores resultados obtidos foram em relação à extroversão, que é a dimensão que apresenta melhores resultados de identificação utilizando os indicadores de redes sociais, conforme verificado na literatura. A pesquisa de Kalghatgi et al. (2015), está baseada na utilização de redes neurais. Foram utilizadas cinco redes neurais, correspondentes às dimensões OCEAN que são treinadas por uma base previamente classificada manualmente. Os autores indicam que o modelo proposto conseguiu realizar a predição com sucesso para 100 usuários, mas não apresenta maiores detalhes sobre acurácia ou outras informações que permitam uma melhor validação da proposta. Este trabalho apresenta uma interessante abordagem utilizando redes neurais que poderia ser melhor aproveitada pela comunidade com a apresentação de resultados mais detalhados que permitissem um comparativo com as demais pesquisas da área. Já o trabalho de pesquisa de Liu et al. (2016) foi utilizada como base de dados o PAN 2015 *Author Profiling Task Dataset* (Rangel et al., 2015), que é uma base obtida a partir de informações de 152 usuários

e suas 14.166 mensagens no *Twitter*. Utilizando uma comparação dos classificadores *Random Forest* e SVM, com um modelo de classificação próprio, o C2W2S4PT (*Character to Word to Sentence for Personality Trait*), os autores demonstraram a melhor performance do modelo proposto, usando o Erro Quadrático Médio (RMS). Este tipo de métrica dificulta a comparação com os demais trabalhos apresentados na área de identificação de personalidade a partir do texto, que utilizam a acurácia como medida principal.

A proposta de Saez et al. (2014) está baseada no desenvolvimento de um modelo para avaliação da personalidade de acordo com o modelo de Eysenck, com coleta de textos da ferramenta *WhatsApp* em língua espanhola. A obtenção das características é realizada com um modelo próprio, ainda em desenvolvimento quando da publicação do artigo, utilizando técnicas de PLN e correlação de palavras e emoções. Os autores realçam as dificuldades de obtenção de bases de referência e outras ferramentas léxicas para a língua espanhola, dificuldades estas presentes também na língua portuguesa. A iniciativa apresenta uma representativa contribuição na identificação de personalidade em língua não inglesa, mas não oferece valores que possam ser usados em comparação com grande parte dos trabalhos descritos nesta seção, que utilizam as bases de língua inglesa.

4.7 Considerações

A partir da análise dos artigos selecionados, foi verificado que a maior parte das propostas de identificação da personalidade, a partir do texto, utilizam como base o idioma inglês, existindo uma lacuna a ser preenchida por iniciativas que verifiquem esta identificação na língua portuguesa, onde também ocorre a carência de bases de referência para treinamento dos classificadores. Como bases de referência, ficou evidenciado que a ESSAYS foi uma iniciativa que alavancou diversos estudos, bem como posteriormente foi disponibilizada a *myPersonality* que contém além do texto, indicadores das atividades no *Facebook* dos autores dos textos que formam a base. Os demais trabalhos verificados utilizaram bases diversas, obtidas em experimentos, ou utilizando informações coletadas em redes sociais. Não foram identificadas pesquisas que utilizaram bases de texto obtidas a partir de atividades educacionais, em português, como fonte de informações para identificação da personalidade. A utilização de uma base de referência, aliado a métricas comuns, permite que os resultados encontrados nos experimentos possam ser comparados entre os diversos modelos. Estas informações nortearam a escolha das bases e métricas utilizadas na presente pesquisa. Em termos dos métodos de representação, foi verificada uma consistente utilização do léxico LIWC, de forma isolada, ou em conjunto com outras técnicas. Foi verificada a consistência dos resultados encontrados quando da utilização do LIWC. Uma segunda categoria de representação está associada às técnicas de PLN, principalmente nas iniciativas mais recentes.

Na Figura 4.8 está indicado o posicionamento da presente pesquisa, identificada como "Buiar (2018)", que utiliza as duas bases de referência mais utilizadas pelas pesquisas investigadas, associando a representação do texto por meio do léxico LIWC e utilizando técnicas de PLN, além de desenvolver um experimento utilizando o idioma português. Foi também obtida uma base própria, em português, utilizada para a validação do modelo desenvolvido, o IP3, obtida a partir das atividades educacionais desenvolvidas em AVA. Com o objetivo de cobrir as lacunas verificadas na literatura, o modelo desenvolvido nesta pesquisa permite a utilização de bases consolidadas, em idioma inglês, para a identificação automática do perfil de personalidade de alunos, baseado somente no texto em português, obtido a partir das atividades educacionais. Oferece ainda a flexibilidade de utilizar diversas técnicas de representação do texto e classificadores, permitindo otimizar o processo de classificação de acordo com cada uma das dimensões do BIG FIVE, e das bases de treinamento disponíveis.

5 Modelo Proposto

Este capítulo apresenta o modelo IP3, desenvolvido durante este projeto de pesquisa. Inicialmente é apresentada a arquitetura geral do modelo, indicando os módulos que o compõem. Na sequência, os módulos de Base de Dados, Representador, Extrator LIWC, Categorizador, Extrator nGRAM, Extrator *Word2Vec* e Classificador são apresentados e descritos, bem como os Parâmetros de Configuração são abordados. O processo de obtenção de uma Base de Validação, em um ambiente educacional brasileiro é descrita na seção seguinte, seguida da descrição do processo de validação do modelo que foi adotada. Ao final são realizadas as considerações sobre o modelo apresentado.

5.1 Modelo IP3

O modelo apresentado por esta pesquisa, o Identificador de Perfil de Personalidade Parametrizável (IP3), consiste em um modelo de aprendizagem de máquina para classificação de texto, como objetivo de realizar o reconhecimento automático dos traços de personalidade entre domínios diferentes. A proposta está baseada no treinamento de um conjunto de classificadores a partir de uma base de dados contendo textos e perfis de personalidade, previamente classificados, e um conjunto de dados formado por textos obtidos a partir das atividades educacionais realizadas em AVA. A Figura 5.1 apresenta a arquitetura deste modelo.

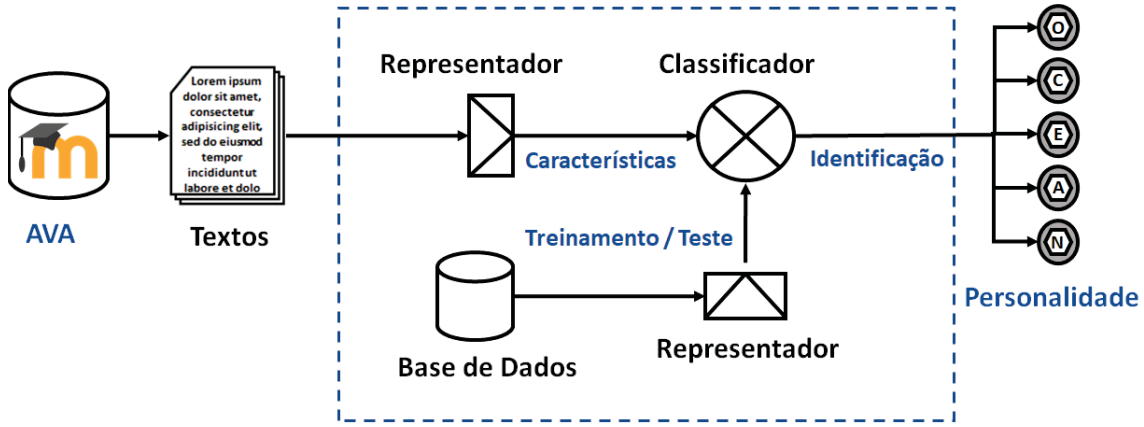


Figura 5.1: Modelo IP3

Este modelo tem como objetivo realizar a identificação das informações de traços de personalidade baseado somente no conteúdo do texto gerado pelos alunos, que no caso do experimento realizado, foi obtido a partir das informações armazenadas nas atividades de fórum de discussão. A princípio este modelo pode ser utilizado com outras fontes que contenham informação textual de autoria dos alunos, oriundas do ambiente AVA ou de outras fontes externas.

As informações oriundas do AVA, são armazenadas em um arquivo no formato “CSV”, sendo associado um identificador para cada autor, que produziu um conjunto de textos, onde todos os parágrafos destes textos são agrupados sem os saltos de linha, em uma célula desta planilha. O modelo foi concebido de forma a permitir que o idioma utilizado na base de treinamento possa ser diferente do utilizado na base a ser identificada, possibilitando um universo maior de opções de bases de treinamento. O modelo proposto por esta pesquisa está utilizando a hipótese simplificadora que considera as dimensões OCEAN independentes entre si, sendo desprezadas as eventuais correlações que possam existir entre estas dimensões. Ou seja, na predição da dimensão *Openness* de um determinado aluno, não estarão sendo consideradas as influências das demais dimensões nesta predição.

5.2 Módulo Base de Dados

Nos trabalhos sobre identificação de personalidade baseado em texto, investigados durante a realização desta pesquisa, foi verificado a separação de uma base previamente rotulada nos dois conjuntos, de treinamento e de teste, para a realização dos testes de validação dos modelos de classificação propostos. Mesmo nos estudos que não utilizaram bases em inglês, os experimentos foram realizados utilizando bases no mesmo idioma. Os estudos que realizaram experimentos somente com textos em língua portuguesa, foram baseados em bases pequenas produzidas pelos autores. Isto se deve em grande a parte pela não existência de bases textuais que contenham classificação do perfil de personalidade dos autores e que estejam disponíveis para uso público.

No modelo apresentado, foi realizada a opção de utilizar uma base consolidada disponível no idioma inglês, para a realização do treinamento e teste do processo de classificação. No experimento realizado foram utilizadas as bases ESSAYS e *myPersonality*. Após a fase de treinamento e testes, com os correspondentes ajustes nos parâmetros do classificador, foi utilizada uma base para validação, com textos em português, denominada de Base UNIVERSIDADE, detalhada na Seção 5.10, com o objetivo de avaliar o modelo proposto em relação aos resultados obtidos com a utilização de bases de treinamento de domínios diferentes.

5.3 Módulo Representador

O módulo “Representador” tem como objetivo obter um conjunto de características que representem as informações que possam inferir os traços de personalidade do autor do texto. Foram utilizadas três técnicas de representação: Léxico, nGRAM e *Word2Vec*. A utilização do léxico permite gerar um conjunto de características baseado nas categorias léxicas das palavras encontradas no texto, a técnica nGRAM permite obter o conjunto de características a partir da frequência de ocorrência dos conjuntos presentes no texto e a utilização de *Word2Vec*, que utiliza a frequência de ocorrência das palavras, mas leva também em consideração o contexto linguístico das palavras.

A Figura 5.2 ilustra como é composto este módulo. Os dados de entrada, presentes em um arquivo no formato “CSV”, contêm as informações sobre um identificador do autor de cada texto, os textos em si e os valores previamente identificados para as classes OCEAN, no caso das bases utilizadas para treinamento, teste e validação. Quando da aplicação deste modelo para a identificação do perfil de personalidade de alunos, após a calibração do classificador, somente as informações do identificador do autor e os textos estarão presentes no arquivo. Na saída serão obtidos os arquivos em formato “SVM”, a serem utilizados pelo módulo “Classificador”.

A partir de um conjunto de parâmetros configuráveis, pode ser selecionado o idioma a ser utilizado, bem como qual o conjunto de características a ser gerado nos arquivos de saída. Esta facilidade permite uma maior flexibilidade na calibração do processo de classificação, bem como ajustes futuros no caso de utilização de outras bases de treinamento.

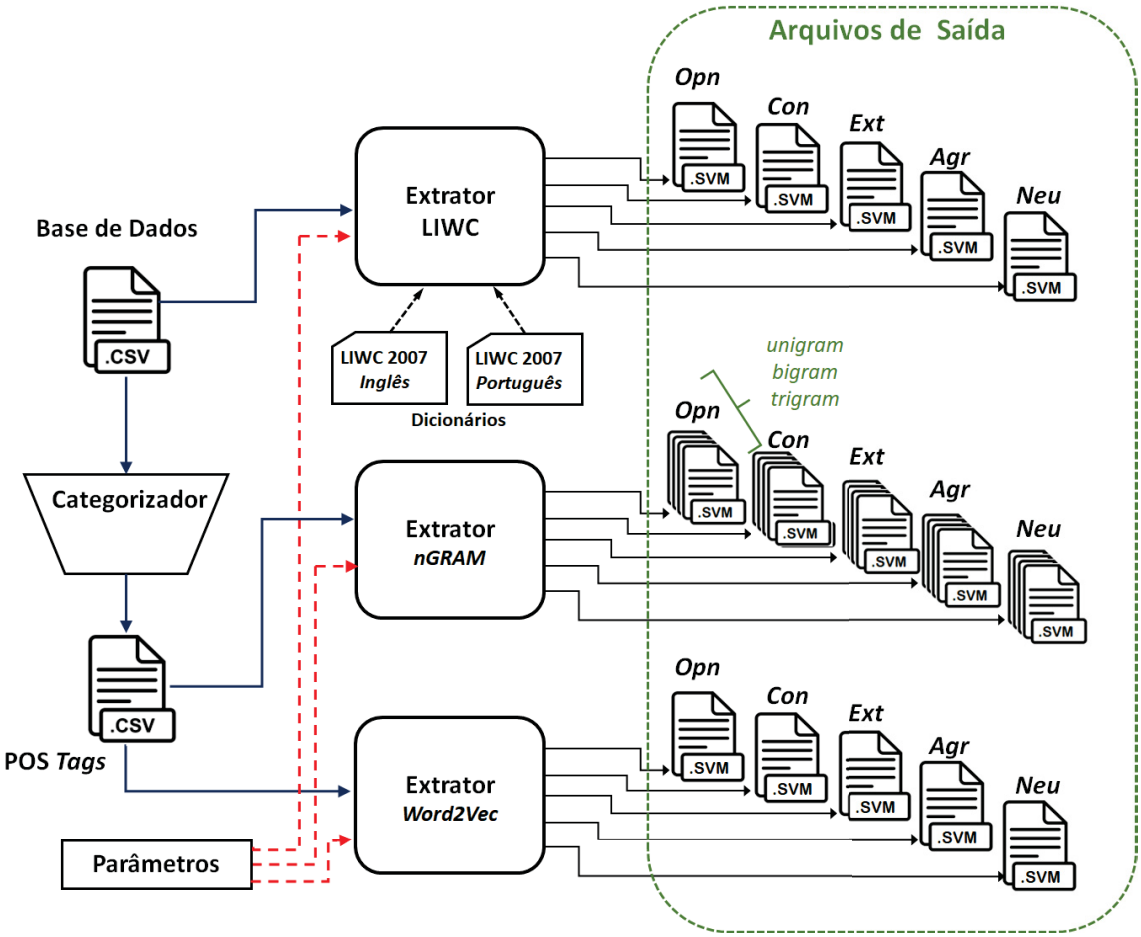


Figura 5.2: Módulo Representador

5.4 Módulo Extrator LIWC

A representação da informação textual através da utilização de um léxico é uma das formas de estruturação da informação textual. O léxico LIWC, apresentado na Seção 2.6.3, pode ser utilizado para obtenção de informações estatísticas de natureza gramática, bem como afetiva. Este léxico foi originalmente desenvolvido para o idioma inglês, nas versões 2001, 2007 e 2015, sendo que a versão 2007 (Pennebaker et al., 2007) foi adaptada para o idioma em português (Balage Filho et al., 2013). A versão do LIWC utilizada nesta pesquisa é a de 2007, que possui as variantes nos idiomas inglês e português, o que permite a realização de um experimento utilizando bases nestes dois idiomas. Para a categorização das palavras presentes no texto, o léxico LIWC utiliza os grupos de categorias:

- Métricas;
- Outras Gramáticas;
- Funcionais;
- Afetivos;

- Sociais;
- Cognitivos;
- Perceptivos;
- Biológicos;
- Necessidades;
- Temporais;
- Relatividade;
- Preocupações Pessoais;
- Linguagem Informal;
- Pontuação.

No experimento realizado foi utilizado o software LIWC2015, que permite selecionar o léxico a ser utilizado, no caso o LIWC 2007 nos idiomas inglês e português, que permite a representação do texto por meio de 80 categorias. Esta ferramenta realiza a leitura de um arquivo no formato "CSV" e gera como saída um arquivo no mesmo formato, indicando a frequência de ocorrência de cada uma das categorias em função de uma sequência de textos fornecida na planilha de entrada. Cada palavra presente no texto em questão é associada a uma ou mais categorias do léxico, para a obtenção da frequência de ocorrência de cada uma das categorias neste texto.

Como resultado do processo de extração é obtido um vetor $V = [v_i, \ldots, v_n]$ que indica a frequência de ocorrência de cada uma das 80 categorias c do léxico L , em cada um dos n textos fornecidos, sendo $v_i = [f_{i,c_1}, \ldots, f_{i,c_{80}}]$. Para a obtenção de cada um dos valores f_{i,c_k}, é obtida a razão entre a contagem de todas as palavras w que pertençam a categoria c_k e que estejam presentes no texto i, pelo número total de palavras n_i presentes no texto i.

$$f_{i,c_k} = \frac{\sum_{j=1}^{n_i} (w_j \in c_k)}{n_i} \tag{5.1}$$

Desta forma, a partir deste vetor V, é gerado um arquivo no formato "SVM", contendo as correspondentes frequências de ocorrência de cada categoria do léxico, identificadas no texto de cada um dos autores. E a partir deste arquivo, são gerados outros cinco arquivos, para cada uma das dimensões OCEAN, contendo o rótulo identificado manualmente do perfil de personalidade de cada autor e as frequências de ocorrência de cada uma das categorias presentes no léxico.

A Figura 5.3 ilustra o armazenamento das frequências de ocorrências obtidas a partir da base ESSAYS utilizando o dicionário LIWC 2007, na forma de uma matriz esparsa, neste caso, para a dimensão *Agreeableness*. Outros quatro arquivos similares são gerados para as outras dimensões do OCEAN, contendo os rótulos do perfil de personalidade especifico, na primeira coluna. Cada um dos arquivos de saída contém uma série de colunas, separados por espaços.

A primeira coluna apresenta a classe identificada, correspondente a classificação manual, para esta dimensão específica. Neste exemplo os valores da coluna "cAGR" são convertidos de "y" e "n", para "1" e "0" respectivamente.

As demais colunas apresentam o armazenamento das frequências de ocorrência, na forma de matriz esparsa, no formato "coluna:valor", onde "coluna" indica um índice para a categoria LIWC correspondente e "valor" o valor encontrado para esta categoria. A categoria 2 (WC) indica o número total de palavras encontradas no texto e a categoria 3 (WPS), indica a quantidade média de palavras por sentença, ambas em números absolutos. As demais colunas indicam os valores de frequência de palavras encontradas na categoria em valores percentuais, como por exemplo, a categoria 4 (Sixltr) indica a frequência de ocorrência de palavras com seis letras ou mais. No caso do autor com Identificador "1997_504851" o valor da categoria 4 indica que 8,76% das palavras encontradas no texto deste autor possuem 6 letras ou mais.

Entrada .CSV

#AUTHID	TEXT	cEXT	cNEU	cAGR	cCON	cOPN
1997_504851	Well, right now I just woke up from a mid-day nap. It's sort of weird, but ever sir	n	y	y	n	y
1997_605191	Well, here we go with the stream of consciousness essay. I used to do things like	n	n	y	n	n
1997_687252	An open keyboard and buttons to push. The thing finally worked and I need not	n	y	n	y	y
1997_568848	I can't believe it! It's really happening! My pulse is racing like mad. So this is wh	y	n	y	y	n
1997_688160	Well, here I go with the good old stream of consciousness assignment again. I fe	y	n	y	n	y
1997_722902	Today. Had to turn the music down. Today I went to the KVRX meeting. I will ho	y	n	y	n	y
1997_724708	Stream of consciousness. What should I write about. Am I supposed to have son	n	n	y	n	n
1997_724794	The RTF305 Usenet site is a piece of garbage! I just sent my first required messa	n	n	n	y	y
1997_628043	I'm really unsure about this assignment because I'm afraid I won't be able to thi	y	y	n	y	y
:::	:::	:::	:::	:::	:::	:::
:::	:::	:::	:::	:::	:::	:::
:::	:::	:::	:::	:::	:::	:::

LIWC

1	2:662	3:0.1742	4:0.0876	5:0.9411	6:0.6495	7:0.1828	8:0.1224	9:0.0967	11:0.006	12:0.0045	13:0.0151	14:0.0604
1	2:647	3:0.0735	4:0.1051	5:0.9629	6:0.6275	7:0.2272	8:0.1484	9:0.1298	10:0.0046	11:0.0046	12:0.0062	13:0.0031
0	2:753	3:0.1177	4:0.1368	5:0.8632	6:0.5618	7:0.1859	8:0.1208	9:0.1102	10:0.004	12:0.0053	13:0.0013	14:0.0651
1	2:345	3:0.0676	4:0.1333	5:0.9449	6:0.629	7:0.2116	8:0.1304	9:0.0841	10:0.0029	11:0.0029	12:0.0377	13:0.0029
1	2:968	3:0.1441	4:0.1333	5:0.9317	6:0.6134	7:0.1894	8:0.1167	9:0.0892	10:0.0077	11:0.0154	13:0.0044	14:0.0727
1	2:717	3:0.0640	4:0.1576	5:0.901	6:0.5802	7:0.1841	8:0.1255	9:0.0907	10:0.0042	11:0.0014	12:0.0167	13:0.0126
1	2:701	3:0.0899	4:0.1484	5:0.9101	6:0.632	7:0.2068	8:0.1084	9:0.0713	10:0.0043	11:0.0185	12:0.0114	13:0.0029
0	2:452	3:0.1159	4:0.1217	5:0.8761	6:0.5553	7:0.1593	8:0.0996	9:0.0796	10:0.0022	12:0.0133	13:0.0044	14:0.0597
0	2:528	3:2.64	4:0.1061	5:0.8845	6:0.6155	7:0.2083	8:0.1458	9:0.0758	10:0.0265	11:0.0019	12:0.0133	13:0.0284
:::	:::	:::	:::	:::	:::	:::	:::	:::	:::	:::	:::	:::
:::	:::	:::	:::	:::	:::	:::	:::	:::	:::	:::	:::	:::

Saída .SVM (*Agreeableness*)

Figura 5.3: Exemplo de Arquivo SVM obtido com LIWC

5.5 Módulo Categorizador

A utilização da técnica de representação estatística da ocorrência das palavras tem sido adotada em modelos de classificação de documentos, como por exemplo, nos sistemas de verificação de autoria de textos. No modelo IP3, proposto neste trabalho, foi adotada uma variação desta técnica, sendo utilizadas informações estatísticas sobre as categorias sintáticas do texto, ao invés da frequência de ocorrência das palavras. Esta proposta está baseada na aplicação de técnicas PLN, conforme ilustrado na Figura 5.4, para a obtenção dos categorias

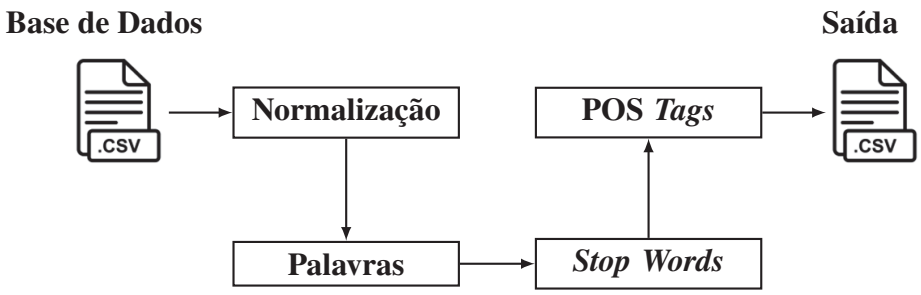


Figura 5.4: Módulo Categorizador

sintáticas das palavras encontradas no texto, utilizando a identificação POS (*part-of-speech tagger*), apresentada na Seção 2.6.5.

O texto fornecido na entrada é submetido a um processo de normalização que realiza a transformação de letras maiúsculas para minúsculas, remoção de caracteres especiais e remoção de tags HTML. A partir disto é realizada a identificação das frases existentes no texto, e as palavras encontradas nestas frases. O vetor de palavras resultantes é submetido à remoção das *Stop Words*. A categorização é realizada utilizando a biblioteca NLTK (*Natural Language Toolkit — NLTK 3.3*) (Bird et al., 2009), em ambiente *Python*. A Figura 5.5 ilustra a obtenção das categorias sintáticas, utilizando o NLTK, para um texto no idioma inglês.

```
>>> import nltk
>>> texto = nltk.word_tokenize("Personality is defined as the set of
    habitual behaviors, cognitions and emotional patterns that evolve from
    biological and environmental factors.".lower())
>>> tags = nltk.pos_tag(texto)
>>> print tags
[('personality', 'NN'), ('is', 'VBZ'), ('defined', 'VBN'), ('as', 'IN'), ('
    the', 'DT'), ('set', 'NN'), ('of', 'IN'), ('habitual', 'JJ'), ('
    behaviors', 'NNS'), (',', ','), ('cognitions', 'NNS'), ('and', 'CC'), ('
    emotional', 'JJ'), ('patterns', 'NNS'), ('.', '.')]
```

Figura 5.5: *Pos Tag* em Inglês

Para a categorização de texto em idioma português, foi necessário o treinamento do *tagger* a partir de um dicionário no idioma português. O dicionário utilizado para a realização do treinamento foi o *corpus* do projeto "Floresta Sintá(c)tica" (Afonso et al., 2002), conforme ilustrado na Figura 5.6.

```
>>> import nltk
>>> import pickle
>>> tagger = pickle.load(open("floresta_treinado.pickle"))
>>> tags = tagger.tag(nltk.word_tokenize("Personalidade é um conjunto de
características que determinam os padrões de pensar, sentir e proceder." .
    lower().decode('utf8')  ))
>>> print tags
[(u'personalidade', u'H+n'), (u'\xe9', u'P+v-fin'), (u'um', u'>N+art'),(u'
    conjunto', u'H+n'), (u'de', u'H+prp'), (u'caracter\xedsticas', u'H+n'),(
    u'que', u'SUBJ+pron-indp'), (u'determinam', u'P+v-fin'), (u'os', u'>N+
    art'), (u'padr\xf5es', u'H+n'), (u'de', u'H+prp'), (u'pensar', u'P+v-inf
    '), (u',', u','),  (u'sentir', u'P+v-inf'),  (u'e', u'CO+conj-c'), (u'
    proceder', u'P+v-inf'), (u'.', u'.')]
```

Figura 5.6: *Pos Tag* em Português

Cada um dos categorizadores dos diferentes idiomas utilizados nos experimentos realizados, o português e o inglês, apresentou uma denominação diferenciada para o conjunto de *tags* que identificam as categorias sintáticas das palavras presentes no texto. Por exemplo, a categoria "substantivo" no *tagger* para inglês é representado como "NN", ao passo que no *tagger* para português é representado como "H+n".

Para contornar esta situação foi realizada a opção de utilizar um formato único de saída, baseado no formato *Universal Part-of-Speech Tagset*, especificado por Petrov et al. (2011), que utiliza as seguinte categorias sintáticas:

- NOUN - substantivos;
- VERB - verbos;
- ADJ - adjetivos;
- ADV - advérbios;
- PRON - pronomes;
- DET - artigos;
- ADP - preposições;
- NUM - numerais;
- CONJ - conjunções;
- PRT - partículas;
- "." - pontuação e
- X - outros.

No caso do categorizador para o idioma inglês, existe disponível a opção " tagset='universal' " conforme exemplo ilustrado na Figura 5.7. Para o categorizador no idioma português foi utilizado uma referência cruzada para a normalização do formato de saída.

```
1 >>> import nltk
2 >>> texto = nltk.word_tokenize("Personality is defined as the set of
      habitual behaviors, cognitions and emotional patterns.".lower())
3 >>> tags = nltk.pos_tag(texto,tagset='universal')
4 >>> print tags
5 [('personality', u'NOUN'), ('is', u'VERB'), ('defined', u'VERB'), ('as', u'
      ADP'), ('the', u'DET'), ('set', u'NOUN'), ('of', u'ADP'), ('habitual', u
      'ADJ'), ('behaviors', u'NOUN'), (',', u'.'), ('cognitions', u'NOUN'), ('
      and', u'CONJ'), ('emotional', u'ADJ'), ('patterns', u'NOUN'), ('.', u
      '.')]
```

Figura 5.7: *Pos Tag* em Inglês com Opção *Universal*

Finalmente o texto é convertido para um vetor que contém a sequencia de categorias semânticas do texto.

$$V_{\text{inglês}} = [\text{'NOUN', 'VERB', 'VERB', 'ADP', 'DET', 'NOUN', 'ADP', 'ADJ', 'NOUN', '.', 'NOUN', 'VERB','DET', 'NOUM', 'ADP'}]$$

$$V_{\text{português}} = [\text{'NOUN', 'VERB', 'DET', 'NOUM', 'ADP', 'NOUN', 'PRON', 'VERB', 'DET', 'NOUN', 'ADP', 'VERB', '.', 'VERB', 'CONJ', 'VERB', '.'}]$$

A Figura 5.8 ilustra a obtenção da saída categorizada, a partir de uma base contendo textos. Neste caso, para cada linha, contendo uma ou mais frases, são obtidas sequencias de *Tags* que correspondem à sequência sintática destas frases. Estas informações serão utilizadas pelos módulos "Extrator nGRAM" e "Extrator *Word2Vec*", descritos nas seções seguintes.

5.6 Módulo Extrator nGRAM

Este módulo tem como finalidade a obtenção das frequências de ocorrências dos nGRAM nos textos apresentados a partir do arquivo gerado pelo módulo "Categorizador", que contém as informações sintáticas do texto (POS *Tags*). Este processo é realizado com utilização da técnica TF-IDF, permitindo a seleção do "*n*" a ser utilizado. Com a utilização do conjunto de categorias sintáticas *universal*, a quantidade máxima de características obtidas pode ir até 12 no caso de *unigram*, 144 para o uso de *bigram* e 1.728 para *trigram*. Estes valores de frequência de

Entrada

#AUTHID	TEXT	cEXT	cNEU	cAGR	cCON	cOPN
1997_504851	Well, right now I just woke up from a mid-day nap. It's sort of weird, but ever sir	n	y	y	n	y
1997_605191	Well, here we go with the stream of consciousness essay. I used to do things like	n	n	y	n	n
1997_687252	An open keyboard and buttons to push. The thing finally worked and I need not	n	y	n	y	y
1997_568848	I can't believe it! It's really happening! My pulse is racing like mad. So this is wh	y	n	y	y	n
1997_688160	Well, here I go with the good old stream of consciousness assignment again. I fe	y	n	y	n	y
1997_722902	Today. Had to turn the music down. Today I went to the KVRX meeting. I will ho	y	n	y	n	y
1997_724708	Stream of consciousness. What should I write about. Am I supposed to have som	n	n	y	n	n
1997_724794	The RTF305 Usenet site is a piece of garbage! I just sent my first required messa	n	n	n	y	y
1997_628043	I'm really unsure about this assignment because I'm afraid I won/t be able to thi	y	y	n	y	y
:::	:::	:::	:::	:::	:::	:::
:::	:::	:::	:::	:::	:::	:::
:::	:::	:::	:::	:::	:::	:::

Pos Tag

1997_504851	ADV . ADV ADV PRON ADV VERB PRT ADP DET ADJ NOUN . PRON VERB NOUN ADP NOUN . CONJ ADV
1997_605191	ADV . ADV PRON VERB ADP DET NOUN ADP NOUN NOUN . PRON VERB PRT VERB NOUN ADP DET ADP
1997_687252	DET ADJ NOUN CONJ NOUN PRT VERB . DET NOUN ADV VERB CONJ PRON VERB ADV VERB NOUN . NO
1997_568848	PRON VERB ADV VERB PRON . PRON VERB ADV ADJ . PRON NOUN VERB VERB ADP NOUN . ADV DET V
1997_688160	ADV . ADV PRON VERB ADP DET ADJ ADJ NOUN ADP ADJ NOUN ADV . PRON VERB ADP PRON VERB AD
1997_722902	NOUN . NOUN PRT VERB DET NOUN PRT . NOUN PRON VERB PRT DET NOUN NOUN . PRON VERB ADV
1997_724708	NOUN ADP NOUN . PRON VERB PRON VERB ADP . NOUN PRON VERB PRT VERB DET NOUN ADP NOUN
1997_724794	DET NOUN NOUN NOUN VERB DET NOUN ADP NOUN . PRON ADV VERB PRON ADJ VERB NOUN . ADV
1997_628043	PRON VERB ADV ADJ ADP DET NOUN ADP PRON VERB ADJ PRON VERB VERB ADJ PRT VERB ADP NOUN
:::	:::
:::	:::

Saída Categorizada

Figura 5.8: Exemplo de Categorização

ocorrência formam o vetor de características utilizado para representar o texto nos processos de classificação.

A Figura 5.9 ilustra um trecho de um arquivo obtido com este módulo, configurado para saída *unigram*. A primeira linha do arquivo, comentada, apresenta as diversas categorias encontradas no texto, e no caso de *bigram* e *trigram*, apresenta as combinações de categorias encontradas, duas a duas e três a três, respectivamente.

```
1 #Source:ADJ:ADP:ADV:CONJ:DET:.:NOUN:NUM:X:PRON:PRT:VERB
2 1 1:0.185146 2:0.260610 3:0.290826 4:0.071850 5:0.185071 6:0.265569
     7:0.434350 8:0.015855 9:0.007688 10:0.389184 11:0.105926 12:0.592982
3 0 1:0.140948 2:0.236333 3:0.245423 4:0.072806 5:0.113622 6:0.319563
     7:0.381769 8:0.019078 9:0.009251 10:0.463765 11:0.077388 12:0.613558
4 1 1:0.138087 2:0.268533 3:0.341314 4:0.067843 5:0.178186 6:0.289900
     7:0.331275 8:0.010535 10:0.454432 11:0.087981 12:0.589770
5 0 1:0.182558 2:0.200288 3:0.244796 4:0.053475 5:0.204738 6:0.245890
     7:0.333813 8:0.009342 10:0.440811 11:0.147116 12:0.658724
6 . . . . . .
```

Figura 5.9: Exemplo de Arquivo SVM contendo Representação *unigram*

O formato de arquivo SVM armazena os dados das frequências de ocorrências, na forma de uma matriz esparsa, sendo que cada célula é representada pelo conjunto "`coluna:valor`". A primeira coluna apresenta o valor da classe atribuída para o vetor de características, neste exemplo, sendo "`1`" ou "`0`".

5.7 Módulo Extrator *Word2Vec*

Este componente tem como objetivo gerar um vetor de características numéricas para representação do texto não estruturado. A partir do arquivo que contém as informações sintáticas do texto (POSTags), gerados pelo módulo "Categorizador", na forma de texto, é obtida uma saída no formato SVM para a realização dos processos de classificação contendo uma quantidade de características específicas.

No experimento descrito no Capítulo 6, estão descritos os resultados obtidos com os diferentes tamanhos de vetores utilizados. No modelo proposto, foi utilizada uma extensão da técnica *Word2Vec* denominada *Doc2Vec* que permite a geração dos vetores de características de um conjunto de palavras, utilizando o modelo CBOW (*Continuous Bag of Words*).

A Figura 5.10 ilustra um trecho de um arquivo SVM, obtido com este módulo, configurado para gerar vetores de características com 10 colunas.

```
1 #Source:d1:d2:d3:d4:d5:d6:d7:d8:d9:d10
2 1 1:1.503385 2:0.496595 3:0.152928 4:0.517205 5:-0.626002 6:0.253716
     7:1.237030 8:-1.061988 9:-1.818225 10:-1.642775
3 0 1:-1.233531 2:-1.218329 3:-0.382708 4:-0.249470 5:0.554006 6:-0.694199
     7:-1.932754 8:1.243911 9:1.340262 10:1.150984
4 1 1:1.803984 2:0.772568 3:0.261776 4:-0.304873 5:-0.589803 6:0.385918
     7:1.383438 8:-1.382511 9:-2.247312 10:-1.648281
5 0 1:0.037831 2:-0.395402 3:0.014730 4:-0.489723 5:-0.663687 6:-0.164174
     7:-1.138368 8:0.468953 9:-0.281916 10:-0.484458
6 1 1:-0.557687 2:-0.026747 3:-0.452190 4:-0.071762 5:0.147437 6:-0.330013
     7:-2.081431 8:0.956106 9:0.076912 10:-0.048509
7 . . . . . .
```

Figura 5.10: Exemplo de Arquivo SVM contendo Representação *Word2Vec*

5.8 Módulo Classificador

A arquitetura proposta para o classificador está baseado no modelo *Ensemble* com a utilização de um conjunto de classificadores específicos para cada uma das dimensões OCEAN. A Figura 5.11 apresenta a estrutura do módulo "Classificador". Para cada uma das dimensões de personalidade, é realizada uma parametrização que indica qual conjunto de dados será utilizado por cada um dos subclassificadores que formam o classificador *Ensemble*, de cada dimensão, bem como quais os algoritmos de classificação que serão utilizados.

Utilizando como exemplo a dimensão *Extraversion*, detalhado na Figura 5.12, pode ser selecionada a base obtida com a aplicação do LIWC na base ESSAYS e o algoritmo GNB para compor o Classificador 1, a base de *bigram* extraída do *myPersonality* com o algoritmo *Random Forest* para compor o Classificador 2, e assim consecutivamente. Os parâmetros configuram a quantidade de classificadores n , os algoritmos e bases b_i , a configuração específica de cada

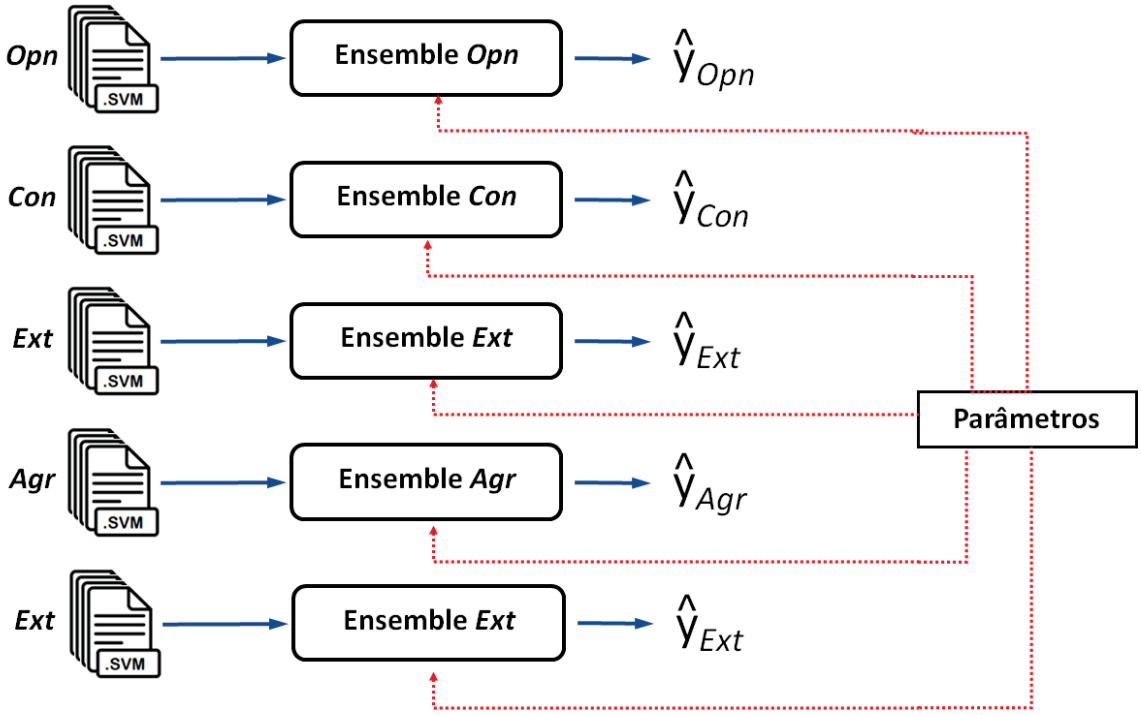


Figura 5.11: Módulo Classificador

classificador e o peso w_i que a estimativa do classificador terá na votação final para definir o valor $\hat{Y}_{dim}$ atribuído para a dimensão em questão, para o autor do texto.

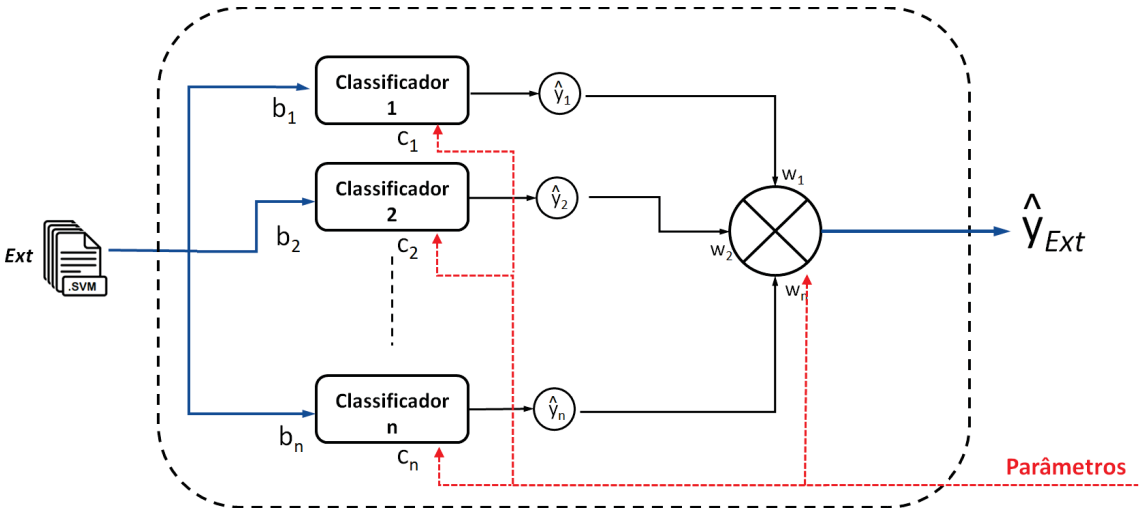


Figura 5.12: Módulo *Ensemble Extraversion*

5.9 Parâmetros de Configuração

A partir de um arquivo de configuração, são especificados os parâmetros a serem utilizados durante a realização do processo de identificação. A Figura 5.13 ilustra o exemplo de arquivo de configuração. As informações estão separadas em blocos de parâmetros. O bloco "executar" contém a lista das tarefas de identificação a serem realizadas. Esta lista descreve um

conjunto de tarefas (*jobs*) que serão executadas na sequência. Para cada tarefa está presente um bloco que deve conter os seguintes parâmetros:

nome - Contém o nome da tarefa, que será visualizada no relatório de saída;

class - Identifica os algoritmos de classificação que serão utilizados para esta tarefa específica;

big5 - Dimensão de personalidade que será identificada.

Um outro conjunto de blocos contém as configurações dos processos de classificação individuais, que devem conter:

class - Identifica o algoritmo de classificação que será utilizado ;

treino - Especifica a base de treinamento a ser utilizada com este classificador;

entrada - Especifica a base que será identificada;

tipo - Define o tipo de representação a ser aplicada, que pode ser *liwc*, *unigram*, *bigram*, *trigram* ou *word2vec*.

```
[executar]
   lista=job1,job2
[bases]
   lista=ess,uni
[ess]
   idioma=en
   arquivo=essays.csv
[uni]
   idioma=pt
   arquivo=universidade.csv
[job1]
   nome=Neuroticism
   class=clf1,clf2,clf3
   big5=neu
[job2]
   nome=Extraversion
   class=clf1,clf2
   big5=ext
```

```
[clf1]
   class=gnb
   treino=ess
   entrada=uni
   tipo=liwc
   peso=1
[clf2]
   class=rforest
   treino=ess
   entrada=uni
   tipo=unigram
   peso=1
[clf3]
   class=knn
   treino=ess
   entrada=uni
   tipo=liwc
   peso=1
```

Figura 5.13: Exemplo de Arquivo de Configuração

5.10 Base de Validação do Modelo IP3

Com o objetivo de obter uma base para validação do modelo proposto, foi realizada um experimento com duas turmas de ensino superior em uma universidade federal. As turmas eram semipresenciais, em que a maior parte das aulas era ofertada por meio da plataforma *Moodle*, na disciplina de Cálculo 2, para alunos de diversos cursos. Apesar das iniciativas junto à turma para que os alunos realizassem atividades, utilizando o fórum de discussão da ferramenta, um número muito pequeno de textos foi obtido, durante os dois meses de realização do experimento. Por meio de investigações junto aos gestores do ambiente de educação à distância, foram identificadas

três turmas do ensino presencial, mas que utilizam o *Moodle* como apoio para realização de atividades no fórum de discussão das disciplinas. Estas turmas, denominadas nesta pesquisa como A, B e C, são de Curso Superior na área de Educação.

Durante a realização da pesquisa, foi obtida uma base de dados contendo textos em idioma português, produzidos por um conjunto de alunos, associados ao perfil de personalidade destes alunos, que estará sendo referenciada como base UNIVERSIDADE. Em uma fase inicial, foi realizada uma breve apresentação dos objetivos da pesquisa aos professores responsáveis, sendo definido que seria realizada a aplicação de um formulário de levantamento de perfil de personalidade a ser preenchido pelos alunos. Este formulário foi baseado no BFI44 (John e Srivastava, 1999) (Seção 2.1.2), adaptado e validado para o idioma português por Andrade (2008), apresentado no Apêndice A. A validação da versão em idioma português foi realizada por Andrade durante o desenvolvimento de sua tese de doutorado no Instituto de Psicologia na Universidade de Brasília, utilizando um universo de mais de 5.000 voluntários das 5 regiões demográficas brasileiras. Na aplicação deste formulário, foi explanado detalhadamente quais eram os objetivos da pesquisa, bem como as características de anonimato e voluntariado da mesma, conforme ilustrado na figura 5.14.

Informações para o(a) participante voluntário(a):

Você está convidado(a) a responder este questionário anônimo que faz parte da coleta de dados da pesquisa "Identificação de Perfil de Personalidade – Big Five", sob responsabilidade do pesquisador Prof. José Antonio Buiar, durante a realização de seu curso de Doutorado no Programa de Pós Graduação em Informática da UFPR – Universidade Federal do Paraná, sob orientação do Prof. Dr. Andrey Ricardo Pimentel.

Caso você concorde em participar da pesquisa, assine a lista abaixo e leia com atenção os seguintes pontos: a) você é livre para, a qualquer momento, recusar-se a responder as perguntas que lhe ocasionem constrangimento de qualquer natureza; b) você pode deixar de participar da pesquisa e não precisa apresentar justificativa para isso; c) sua identidade será mantida em sigilo; d) caso você queira, poderá ser informado(a) de todos os resultados obtidos com a pesquisa, independentemente do fato de mudar seu consentimento em participar da mesma.

Figura 5.14: Informação para Participação Voluntária em Pesquisa

Após a coleta dos formulários preenchidos pelos alunos, foi realizada a compilação dos resultados para levantamento do perfil de personalidade, correspondendo aos valores entre 1 e 5 para cada uma das cinco dimensões do BIG FIVE. Foram respondidos 10 formulários de resposta na turma A, 22 na turma B e 17 na turma C, com um total de 49 formulários. Destes alunos, 4 não tinham registrado atividades de fórum no *Moodle*, sendo então descartados da base, que ficou com os textos de 45 alunos. Além destes indicadores, foram extraídos os conteúdos das mensagens postadas por estes alunos no Moodle, nas atividade de fórum. Este procedimento está ilustrado na Figura 5.15.

5.11 Validação do Modelo

A partir do modelo IP3 apresentado neste capítulo, foi realizado um processo de validação deste modelo, que está apresentado no Capítulo 6. Esta validação realizou uma série de ensaios utilizando as bases de treinamento ESSAYS e *myPersonality* sobre um conjunto de classificadores, utilizando as diversas técnicas de representação propostas. Ao final, é realizado um experimento de identificação automática do perfil de personalidade dos alunos, a partir dos

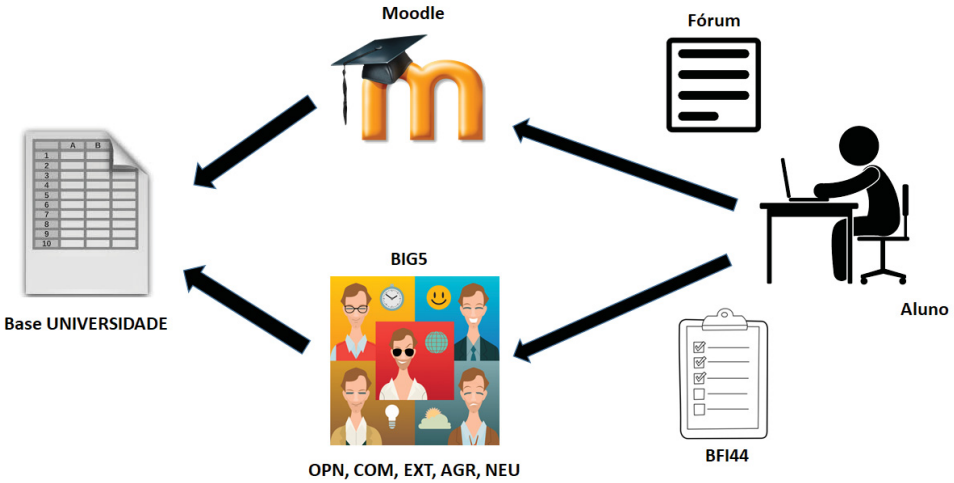


Figura 5.15: Base Universidade

textos presentes na base UNIVERSIDADE e verificados os resultados obtidos com diversas configurações do modelo IP3.

5.12 Considerações

O objetivo do modelo proposto nesta pesquisa é apresentar uma alternativa automática para identificação do perfil de personalidade de alunos, a partir da inferência dos traços de personalidade presentes nos textos por estes redigidos, nas atividades educacionais. Este processo de identificação está baseado na teoria dos traços de personalidade e na sua abordagem léxica (Allport, 1937; Murray, 1938; Cattell, 1957; Goldberg, 1990; Norman, 1963) amplamente utilizada como referência nos processos de identificação manual dos traços de personalidade na área de Psicologia. Conforme verificado por Goldberg (1981), as características de personalidade não podem ser medidas com precisão.

Mesmo com a utilização de técnicas consolidadas de classificação, as iniciativas de identificação de personalidade devem levar em conta as limitações deste processo. Um ponto crítico da identificação da personalidade diz respeito à manifestação dos traços de personalidade nos textos produzidos, e no caso dos textos obtidos a partir de atividades educacionais, a manifestação destes traços tendem a ser mais restritas. Nos experimentos verificados na literatura, os indivíduos manifestaram livremente os traços de personalidade nos textos produzidos em redes sociais ou em redações livres. Já no caso do texto oriundo de atividades educacionais, o aluno registra no texto uma quantidade relevante de informações referentes ao objetivo da atividade, o que poderia dificultar a identificação dos traços reais da personalidade.

Outro ponto que deve ser levado em consideração no processo de identificação que utiliza aprendizado de máquina é o tamanho das bases de treinamento. No caso de bases previamente classificadas com perfil de personalidade, não estão disponíveis bases com grandes quantidades de informações que permitam aumentar a precisão do processo. Esta limitação foi observada nos ensaios realizados utilizando o IP3. O modelo proposto foi concebido com a possibilidade de utilização de bases diversas, inclusive em idiomas distintos, permitindo minimizar esta limitação.

6 Experimentos e Resultados

Este capítulo descreve os experimentos utilizados, bem como os resultados obtidos. Inicialmente é apresentada a metodologia adotada e as metas atingidas. A seção seguinte apresenta os materiais, onde estão relacionadas as bases de dados, os classificadores e as ferramentas de *software* utilizadas. Na sequência do capítulo, são apresentados os resultados da verificação inicial realizada com as bases de dados. As seções seguintes demonstram os resultados obtidos com os testes de classificação utilizando as bases de treinamento ESSAYS e *myPersonality*, com as técnicas de representação por meio do léxico LIWC e das técnicas de nGRAM e *Word2Vec*. Na sequência são apresentados os resultados obtidos na utilização do modelo IP3 para a identificação do perfil de personalidade de alunos, com a utilização da base de validação UNIVERSIDADE. Os resultados parciais desta pesquisa, na forma das publicações geradas durante o desenvolvimento da tese são então apresentados e ao final, são descritas as considerações sobre os resultados obtidos.

6.1 Metodologia

O passo inicial do desenvolvimento do experimento foi a definição das metas a serem atingidas, com o objetivo de validar o modelo desenvolvido durante esta pesquisa. A partir destas metas, foram selecionais os materiais a serem utilizados, que compreendem as bases de treinamento, os programas desenvolvidos especificamente para estes experimentos, as ferramentas e programas de uso geral que foram utilizado, bem como os classificadores e as ferramentas relacionadas.

Com o ambiente de testes definido, foram realizados inicialmente ensaios gerais com as bases de dados, para obter a sua composição, informações estatísticas sobre o conteúdo dos dados e distribuição das classes. Para isto foram desenvolvidos programas em *Python* utilizando bibliotecas de PLN. Estes ensaios possibilitaram a obtenção de informações qualitativas e quantitativas das bases ESSAYS, *myPersonality* e UNIVERSIDADE.

Na etapa seguinte foram avaliados os modelos de representação LIWC, nGRAM e *Word2Vec*, aplicados de forma isolada sobre as bases ESSAYS e *myPersonality*. Inicialmente foi feita a seleção de um conjunto de classificadores e por meio de um conjunto de programas desenvolvidos em *Python*, foram realizados testes de verificação da acurácia obtida com a representação LIWC nos textos destas bases. A obtenção dos arquivos secundários para a realização dos testes de classificação foram obtidos a partir do módulo "Extrator LIWC" descrito na Seção 5.4. Na sequência, por meio da utilização do módulo "Categorizador", apresentado na Seção 5.5, foram obtidos os arquivos contendo a representação das categorias sintáticas do texto das bases ESSAYS e *myPersonality*.

Após este processo foram obtidos os arquivos secundários contendo a representação nGRAM, de forma isolada, nas duas bases, utilizando o módulo "Extrator nGRAM", apresentado na Seção 5.6. Com o desenvolvimento de um conjunto de programas na linguagem *Python*, foram

realizados então os ensaios utilizando a representação nGRAM nas modalidades *unigram*, *bigram* e *trigram*, apresentando os resultados comparativos obtidos com os diversos classificadores, para cada uma das dimensões OCEAN, em cada uma das bases.

De forma similar, foi realizado um ensaio para a verificação dos resultados obtidos com representação *Word2Vec* dos textos destas duas bases. Os arquivos secundários foram obtidos com a utilização do módulo "Extrator *Word2Vec*", descrito na seção 5.7, sendo também apresentados os resultados comparativos da acurácia obtida com esta forma de representação do texto, aplicado às dimensões OCEAN em cada uma das duas bases. Neste ensaio também foram comparados os resultados obtidos com a utilização de diversos tamanhos de vetores de características *Word2Vec*, na faixa de 10 a 500 características.

Com base nos resultados observados nos ensaios realizados com as bases ESSAYS e *myPersonality*, foram estabelecidos os valores de referência para a realização dos ajustes e validação do modelo IP3, de acordo com as duas bases de treinamento, os diversos classificadores utilizados, bem como as formas de representação LIWC, nGRAM e *Word2Vec*, combinadas de diversas formas.

Foram então realizados os ensaios com a base UNIVERSIDADE, contendo textos que os alunos inseriram nas atividades de fórum em AVA, conforme o processo descrito na Seção 5.10. Estes ensaios compararam os valores obtidos com a utilização do modelo IP3 na identificação do perfil de personalidade dos alunos, a partir dos textos registrados na base UNIVERSIDADE, utilizando diversas combinações de forma de representação e classificadores, para cada uma das dimensões OCEAN. Para isto, também foram obtidos arquivos secundários contendo a representação do texto dos alunos de acordo com as técnicas LIWC, nGRAM e *Word2Vec*.

A utilização da identificação da personalidade, a partir dos textos em idioma português, presentes na base UNIVERSIDADE, utilizando como referência os textos em inglês, presentes nas bases ESSAYS e *myPersonality*, foi possibilitado pela utilização dos dicionários LIWC nos idiomas português e inglês, presentes na versão 2007 do LIWC, bem como pela definição da utilização da representação das categorias do texto com o *Pos Tag* do tipo "*universal*", conforme descrito na Seção 5.5. Ao final são apresentados os resultados obtidos para cada uma destas dimensões do modelo BIG FIVE.

6.2 Metas

Com base na hipótese de que seria possível a identificação do perfil de personalidade baseado nos textos desenvolvidos em idioma português pelos alunos durante a realização de atividades educacionais, foi desenvolvida nesta tese o modelo IP3, para verificação da validade da hipótese, por meio da realização de um conjunto de experimentos, realizados com dados obtidos em um ambiente educacional. Os experimentos realizados permitiram avaliar os resultados obtidos com o processo de identificação do perfil de personalidade de alunos, a partir de uma base contendo dados textuais das atividades escolares realizadas em AVA, comparando com os valores obtidos por meio de um levantamento manual.

6.3 Materiais

Esta seção apresenta os materiais utilizados no experimento, descrevendo as bases de dados, os classificadores e as ferramentas de *software* utilizados.

6.3.1 Bases de Dados

Uma das bases utilizadas no experimento é a base ESSAYS, que foi obtida por Pennebaker e King (1999), apresentada na Seção 2.5.1, na qual 2.467 voluntários escreveram uma redação por cerca de 20 minutos. A classificação manual foi realizada com a aplicação do questionário BFI-44. A outra base utilizada no experimento é a *myPersonality*, descrita na Seção 2.5.2, foi obtida a partir da coleta de textos do *Facebook*, de voluntários que participaram do projeto homônimo *myPersonality* (Kosinski et al., 2015). Estas duas bases, com textos no idioma inglês, foram utilizadas nos testes iniciais dos classificadores e verificação das formas de representação de texto, bem como utilizadas como bases de treinamento para o modelo IP3.

Além destas duas bases públicas, durante a condução da presente pesquisa, foi obtida uma base própria, em português, obtida a partir de informações de textos oriundos de atividades educacionais, especificamente de fórum em AVA, denominada base UNIVERSIDADE. A classificação manual do perfil de personalidade dos alunos foi obtida com a aplicação do formulário BFI-44 em português, conforme descrito nas Seção 5.10. Esta base foi utilizada para a validação do modelo IP3, utilizando dados obtidos em um ambiente de ensino universitário no Brasil.

6.3.2 Processamento de Linguagem Natural

Para a realização das tarefas de Processamento de Linguagem Natural (PLN) durante o experimento, foi utilizada a biblioteca *Natural Language Toolkit — NLTK 3.3* (Bird et al., 2009). Com a utilização de um conjunto de programas desenvolvidos durante a condução da presente pequisa, utilizando a linguagem *Python* em ambiente *Linux*, foram realizadas as tarefas de PLN necessárias para a preparação dos textos oriundos das bases, utilizadas nos processos de classificação.

6.3.3 Classificadores

Os algoritmos de classificação utilizados nos experimentos foram obtidos utilizando a biblioteca *scikit-learn Machine Learning in Python* (Pedregosa et al., 2011) (SCIKIT), sendo referenciados nos resultados apresentados como:

BASELINE - classificador *DummyClassifier*, utilizado como referência inicial para os resultados de classificação;

GNB - classificador *GaussianNB* que implementa o algoritmo *Naïve Bayes*;

kNN - classificador *Nearest Neighbors* utilizando como predição os vizinhos mais próximos, sendo adotado o número de vizinhos igual a “3”;

LR - classificador *LogisticRegression* que utiliza o modelo linear de regressão *Logistic*, com resolvedor *liblinear*;

MLPC - classificador *Multi-layer Perceptron*, utilizando um resolvedor *lbfgs*, com 5 camadas de 2 neurônios;

RFOREST - classificador *RandomForestClassifier*, utilizando um conjunto de classificadores do tipo *Decision Tree*, sendo adotado o número de arvores como “10”;

SVM - classificador *Support Vector Machines*, do tipo *C-Support Vector*, utilizando um resolvedor *lbfgs*, com *kernel* do tipo "rbf".

Os ensaios realizados nas bases para verificação da acurácia, utilizaram a separação das bases de dados nos conjuntos de treinamento e de teste, de acordo com o método de validação cruzada *k-Fold*, sendo adotado o valor de "k=5". Neste processo, as bases secundárias obtidas, foram separadas com estratificação, ou seja, foram mantidas as proporções das classes da base original nas bases secundárias, em cada uma das dimensões OCEAN. Os valores da acurácia média obtida estão indicados como $\overline{acc}$, sendo que o desvio padrão também é indicado ao lado.

6.3.4 Ferramenta LIWC2015

Para a realização dos testes de classificação é necessária a preparação das bases de dados, que contêm os textos e a informação sobre o perfil BIG FIVE, obtido previamente pela classificação manual, para serem processados pelos programas desenvolvidos durante o experimento. No processo de avaliação dos classificadores, será necessário confrontar os valores BIG FIVE obtidos pelo classificador com os valores originais, obtidos pela classificação manual, que estão presentes nestas bases de dados. O classificador, por sua vez, necessitará dos vetores de características que representam cada texto, para realizar a estimativa dos valores BIG FIVE.

Com a utilização da ferramenta LIWC2015, ilustrada na Figura 6.1, foram gerados arquivos secundários, no formato "SVM", a partir dos arquivos "CSV", correspondentes às bases de dados de entrada. Estes arquivos secundários contém os vetores de caraterísticas que

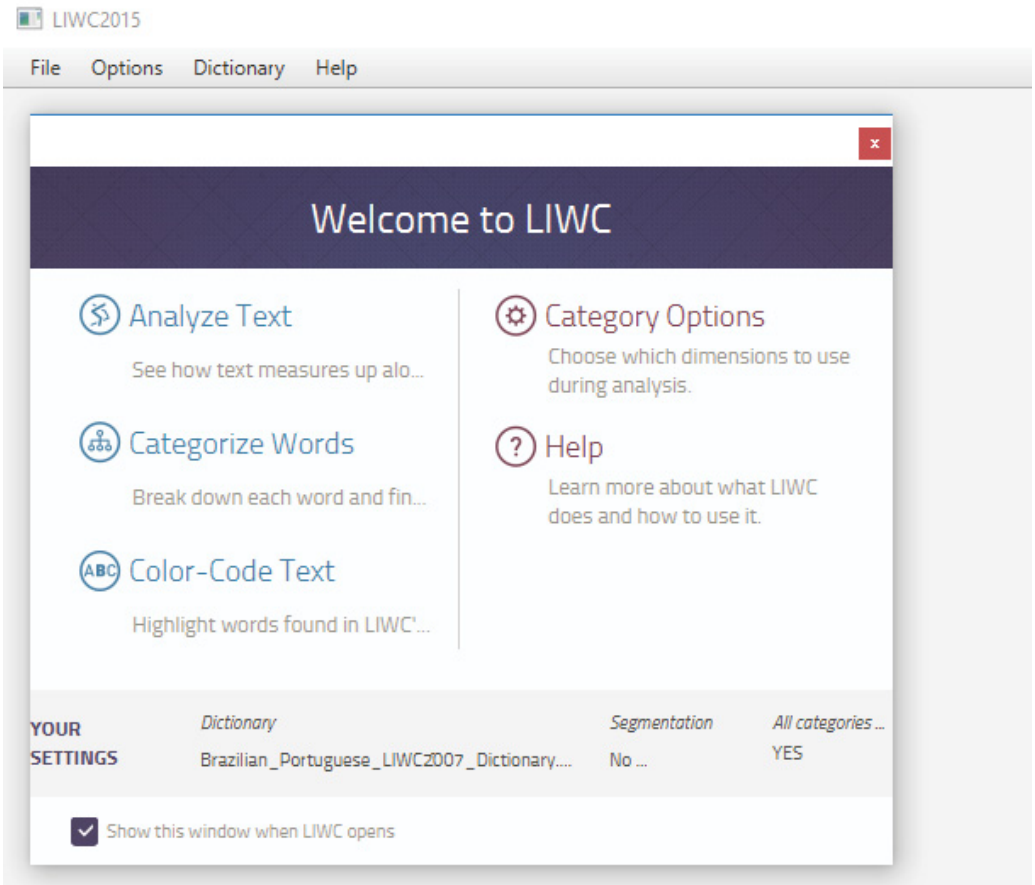


Figura 6.1: Ferramenta LIWC2015

representam cada texto, com a forma de representação LIWC. Para cada uma das bases utilizadas, foram gerados cinco arquivos, correspondentes a cada uma das dimensões OCEAN. Desta forma estará disponível um conjunto de arquivos que podem ser utilizados para o treinamento dos classificadores em cada uma das dimensões OCEAN para cada uma das bases utilizadas.

Para o caso das outras formas de representação utilizadas, nGRAM e *Word2Vec*, foram utilizados programas desenvolvidos no decorrer da presente pesquisa. Estes programas foram codificados em linguagem *Python* e executados em ambiente *Linux*. Desta forma foram obtidas as bases secundárias para utilização nos experimentos, contendo os vetores de características obtidas com estas outras técnicas de representação de texto. Estes arquivos também foram gerados no formato *SVM*, sendo obtido um arquivo para cada uma das dimensões OCEAN correspondentes, conforme ilustrado na Figura 5.2, apresentada no Capítulo 5.

6.4 Verificação das Bases de Dados

Esta seção descreve os experimentos realizados com as bases de dados utilizadas, a ESSAYS e a *myPersonality*, apresentadas na Seção 2.5 e a UNIVERSIDADE, descrita na Seção 5.10. O objetivo destes ensaios é a verificação do formato de apresentação das informações nestas bases, a distribuição dos dados, bem como de outras informações relevantes para esta pesquisa.

6.4.1 Base ESSAYS

Os textos apresentados nesta base são, em geral, bem estruturados gramaticalmente, apresentando uma linguagem formal, que foram obtidos a partir de redações propostas pelos pesquisadores (Pennebaker e King, 1999), para os voluntários que participaram da pesquisa de onde originou esta base. Um exemplo de um trecho, obtido a partir de um dos textos registrados nesta base, é apresentado na sequência.

> " *Well, I feel good about the fact that I am getting this assignment done well before it is due. Today is one of those days that I feel really motivated to do my homework, as opposed to those days in which I don't do anything worthwhile. The excitement of college is starting to wear off and I think that the reality of the fact that I am here is finally sinking in. I really hate the way this typing field doesn't automatically move the sentence down to the next line! I really don't seem to be thinking about anything interesting right now. I am just feeling average, not extremely excited or unhappy. I really cannot think of anything to type. I think my mind is clearing itself like it usually does when I sit down to right a paper. No stray thoughts seem to be coming to me. I am fairly excited about this psychology course. I think this course will not only be very interesting but helpful as well because I plan to go into medicine. Boy, this twenty minutes is going by slowly. I think I might be typing too much too fast. Perhaps I am supposed to sit and wait till a thought comes to me before I type. I have tried to type in my current thoughts and feelings. My roommate is typing on his computer as well, annoying. Now he has turned on his fan, which is fairly loud . he switched it off. Still no stray thoughts. I guess composing these sentences are thoughts. This assignment is all I am thinking about right now.* "

A estrutura da base ESSAYS é apresentada na Tabela 6.1. Nesta representação, a coluna "Authid" indica uma referência ao autor do texto, mantendo o anonimato de seus verdadeiros

nomes, bem como uma referência ao ano em que foi realizada a coleta do texto de cada usuário. A coluna "Text" apresenta o texto obtido com cada autor, em idioma inglês, sendo que os parágrafos do texto em questão, foram apresentados em uma mesma linha. As colunas de "C_{Ext}" até "C_{Opn}" apresentam a identificação de cada uma das cinco dimensões OCEAN, sendo que "Y" indica a presença da dimensão e "N" indica a ausência.

Tabela 6.1: Formato da Base ESSAYS

Authid	**Text**	C_{Ext}	C_{Neu}	C_{Agr}	C_{Con}	C_{Opn}
1997_504851	Well, right now I just woke up...	n	y	y	n	y
1997_605191	Well, here we go with the stre...	n	n	y	n	n
1997_687252	An open keyboard and buttons t...	n	y	n	y	y
1997_568848	I can't believe it! It's real...	y	n	y	y	n
1997_688160	Well, here I go with the good ...	y	n	y	n	y
1997_722902	Today. Had to turn the music d...	y	n	y	n	y
1997_724708	Stream of consciousness. What ...	n	n	y	n	n
:	:	:	:	:	:	:
2004_498	Man this week has been hellish...	n	y	n	n	y
2004_499	I have just gotten off the pho...	n	y	y	n	y

Inicialmente foram verificados os dados estatísticos sobre o conteúdo do texto armazenado nas bases, utilizando as técnicas de PLN da biblioteca NLTK. Os resultados desta verificação estão apresentados na Tabela 6.2.

Tabela 6.2: Características da Base ESSAYS

Característica	**Valor**
Idioma	Inglês
Linhas no Arquivo	2.467
Total de Usuários	2.467
Total de Palavras	1.833.243
Quantidade Média de Palavras por Usuário	743
Palavras Diferentes	38.679
Quantidade Média de Ocorrências por Palavra	47
Quantidade de Sentenças	120.515
Quantidade Média de Palavras por Sentença	15

Não foi identificado um consenso, na literatura investigada, sobre a quantidade mínima de palavras necessárias para identificação da personalidade, mas sim, o quanto maior esta quantidade mais precisos serão os resultados obtidos.

Os indicadores do número total de palavras e da quantidade média de palavras pode ser usado como comparativo, quando da utilização de bases diversas de classificadores, para verificação da correlação entre este parâmetro e a acurácia obtida nos ensaios. De acordo com Wylie (2014), uma quantidade de 14 palavras por sentença, em textos no idioma inglês, permite a compreensão de mais de noventa por cento da informação por parte dos leitores, e esta razão diminui com o aumento da quantidade de palavras. A característica verificada nesta base, onde a quantidade média verificada foi de 15 palavras por sentenças, indica que os textos obtidos pelos pesquisadores que compilaram a base ESSAYS, estão dentro destes padrões. Outro parâmetro que demonstra a abrangência desta base é a presença de mais de 38 mil palavras diferentes.

Outra análise realizada, foi a frequência de distribuição das classes em cada uma das dimensões BIG FIVE, conforme ilustrado na Figura 6.2. Os dados apresentados por esta base

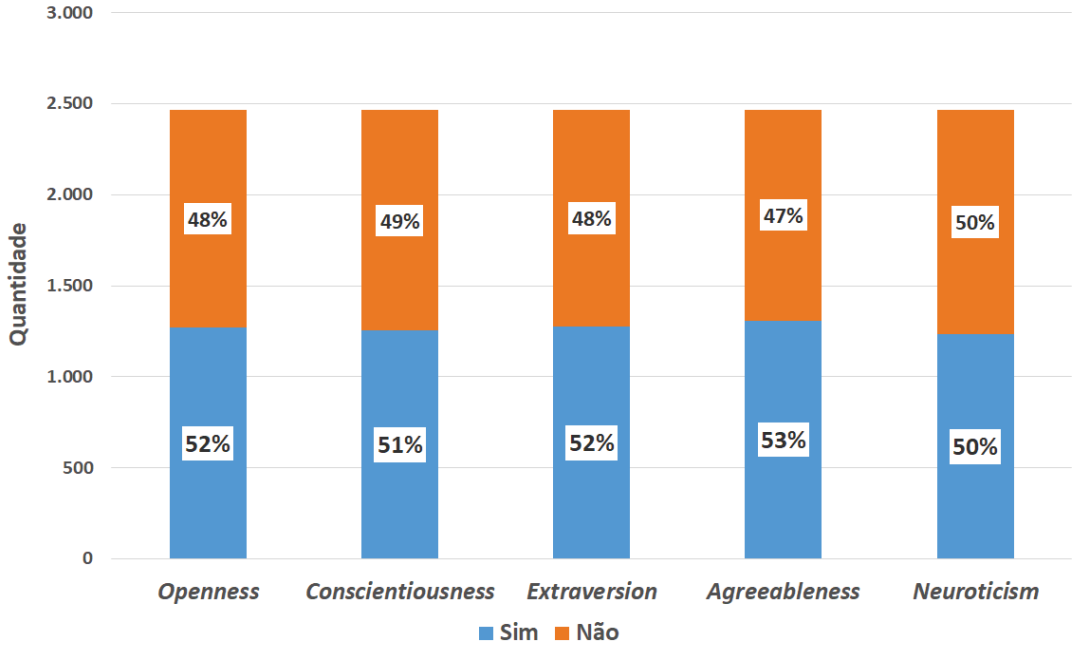


Figura 6.2: Distribuição de Classes na Base ESSAYS

possuem uma distribuição balanceada entre a presença e ausência de cada um dos fatores OCEAN. Uma base que tenha a distribuição das classes balanceada, ou seja, com a presença de cerca da metade das ocorrências pertencendo a cada uma das classes, tende a refletir melhores resultados nos processos de classificação binária (He e Garcia, 2008).

6.4.2 Base *myPersonality*

Os textos presentes nesta base, foram obtidos a partir dos *posts* registrados no *Facebook*, pelos voluntários, e são de natureza livre, não havendo um grande formalismo, apresentando gírias e contrações de palavras. Um exemplo de texto encontrado nesta base é apresentado a seguir, em sua íntegra.

> *"<3 scotch tasting golfing! fore! bruno anyone? <3 boarding at the beach :)! Kitten Mitten! Hot pot! "Holy Haruhi, season 2's out?!"A recent online exchange: Other guy: "I have to imagine it's wholly possible that someone could understand humans so well, yet be completely unable to read, write, or do math. In today's society, people like that would be shunned. Veritas If you can't r has been accepted into grad school. Someone tell me I'm not insane, please. passed his F//M exam. Huzzah! is in Ottawa. Illinois, not Canada. There goes my GPA. "40 years ago, Armstrong said 'One Small Step for Man'. When can one of us go back there and shout, 'Honey, I'm home!' or something?"is going to go on a 1000 mile trip to Iowa City tomorrow and Friday. This is going to be a long drive. has just finished his first night at reffing. Much to improve on. lit his fried chicken on fire and now has more stuff to clean. Great. Wishes he had an actual bed (or even a mattress). It'd make dealing with these suspiciously-flu-like symptoms easier. Recidivism (n.)*

- 1. Committing new offenses after being punished for a crime. 2. Chronic repetition of criminal or other antisocial behavior. Replace "criminal"with "stupid", and I think I've found my problem."

A Tabela 6.3 ilustra a estrutura reduzida desta base, que agrega os textos produzidos por 250 usuários voluntários, que são apresentados em 9917 registros. Cada registro corresponde a

Tabela 6.3: Formato da Base *myPersonality*

Authid	**Status**	$\mathbf{C_{Ext}}$	$\mathbf{C_{Neu}}$	$\mathbf{C_{Agr}}$	$\mathbf{C_{Con}}$	$\mathbf{C_{Opn}}$
b7b...946c4	likes the sound of thunder....	n	y	n	n	y
b7b...946c4	likes how the day sounds in this new son...	n	y	n	n	y
:	:	:	:	:	:	:
318...619c0	has bed bugs..... ewwwww!...	y	y	n	y	n
318...619c0	and mosquito bites...	y	y	n	y	n
:	:	:	:	:	:	:
318...619c0	Ten Movies to Watch Right Now (and some ...	y	y	n	y	n
ecb...061c7	is stuck on Band-Aid brand, cuz Band-Aid...	y	n	y	n	y
:	:	:	:	:	:	:
ea2...3c8ba	is studying hard for the G.R.E....	y	y	y	y	y
553...a9d2d	snipers get more head...	n	y	n	n	y
a28...31e1e	Last night was amazing! Not only did I s...	y	y	n	y	y

um *post* realizado por um determinado usuário, ocorrendo assim a apresentação de diversos *posts* por usuário. A coluna AUTHID corresponde ao identificador único de cada usuário, mantendo seu anonimato. A coluna STATUS indica o texto propriamente dito. As colunas C_{Ext} a C_{Opn}, por sua vez, apresentam estas dimensões representadas como "Y" e "N", tal como na base ESSAYS.

Além destas colunas indicadas na Tabela 6.3, esta base possui outras colunas que indicam a data e a hora da realização do *post* e características sobre o perfil social, como o número de amigos na rede. Estas informações sociais não estão sendo utilizadas no escopo da presente pesquisa, que tem a premissa de verificar a identificação dos traços de personalidade somente utilizando as informações presente no texto.

O levantamento dos dados estatísticos desta base foi realizado, sendo apresentado na Tabela 6.4. Uma característica inicialmente observada é a de que apesar de possuir quase

Tabela 6.4: Características da Base *myPersonality*

Característica	**Valor**
Idioma	Inglês
Linhas no Arquivo	9.917
Total de Usuários	250
Total de Palavras	177.689
Quantidade Média de Palavras por Usuário	710
Palavras Diferentes	20.254
Quantidade Média de Ocorrências por Palavra	8
Quantidade de Sentenças	17.138
Quantidade Média de Palavras por Sentença	10

quatro vezes mais registros do que a base ESSAYS, esta base somente apresenta os dados de

250 voluntários, contra os quase 2500 da base ESSAYS. Isto também reflete no número total de palavras, que é em torno de 10% da quantidade presente na base ESSAYS. Mas a quantidade média de palavras das duas bases está na mesma ordem de grandeza. Outra característica desta base, que foi obtida a partir de registros no *Facebook*, é a presença de frases mais curtas, com uma média de 10 palavras por sentença.

Em face da característica da base *myPersonality* possuir vários registros para cada um dos voluntários, foi levantada a hipótese de qual forma seria mais adequada para utilizar esta base nos ensaios: agrupando os diversos registros de cada voluntário em um registro único formando uma base com 250 registros ou permanecendo com a base original de 9.917 registros. Inicialmente foi realizada uma verificação nos estudos apresentados no Capítulo 4, sendo constatado que a maioria dos autores utilizou a base em sua forma original.

Apesar disto, na presente pesquisa, foram realizados experimentos de classificação utilizando a base original e a base com agrupamento dos *posts*, sendo que não foram verificadas diferenças significativas nos resultados, sendo portanto, realizada a opção metodológica de utilizar a base *myPersonality* em seu formato original, sem agrupamento dos *posts* de um mesmo usuário, e também utilizando somente as informações contidas no texto, não utilizando os registros referentes a informações de rede social, presentes nesta base.

Os dados apresentados por esta base possuem a distribuição entre as classes apresentada na Figura 6.3. Pode ser verificado que esta base não possui o balanceamento similar à base ESSAYS. A dimensão *Openness*, por outro lado, apresenta a predominância de 74%, e a dimensão *Neuroticism* ocorre em somente 37% dos participantes.

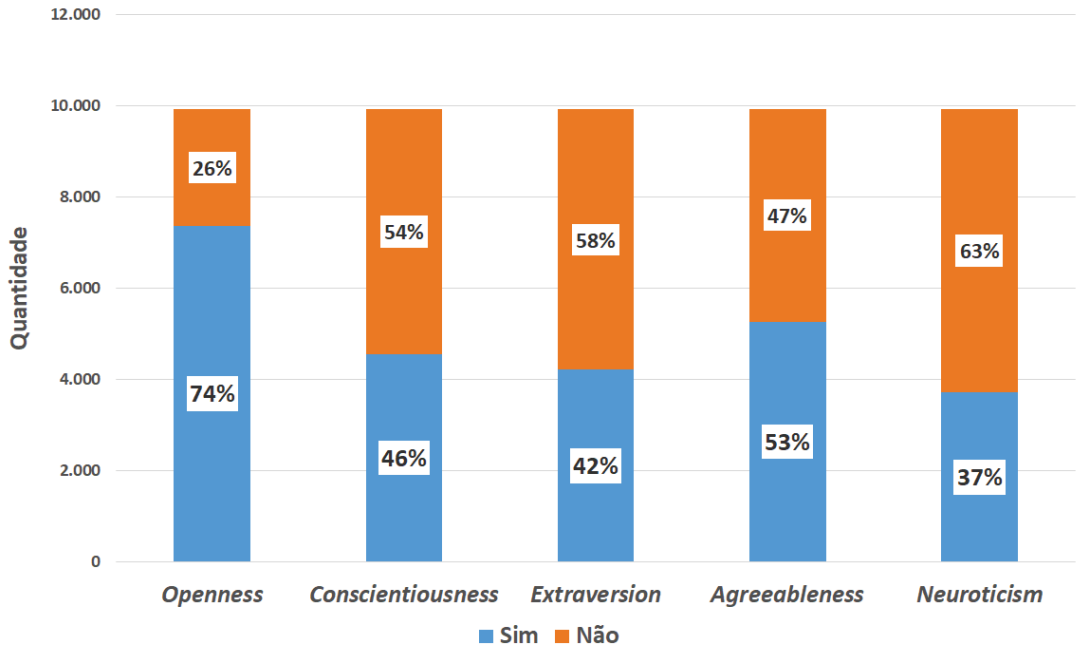


Figura 6.3: Distribuição de Classes na Base *myPersonality*

Nos processos de classificação binária, que foram realizados nos ensaios descritos na presente pesquisa, quando da separação da base em conjuntos de treinamento e teste, foram mantidas as proporções de balanceamento da base original, para minimizar o efeito de balanceamento nos testes realizados. De qualquer maneira, deve-se ter cautela em relação aos resultados apresentados, especificamente para estas dimensões onde o desbalanceamento é mais acentuado.

6.4.3 Base UNIVERSIDADE

Esta base foi obtida a partir de textos gerados por alunos quando da realização de atividades educacionais, em AVA. Além dos textos, esta base contém os valores OCEAN que foram obtidos com a identificação manual realizada por meio da aplicação do questionário BFI44, conforme descrito na Seção 5.10. Estes textos apresentam razoável formalismo em sua estrutura e apresentam uma característica mais descritiva e menos pessoal, do que os textos presentes nas bases de treinamento, já abordadas. Um exemplo de texto coletado nesta base é apresentado a seguir. O texto original registrado pelos alunos foi mantido, sem correções ortográficas ou gramaticais.

> " *Arizona/USA: Desde 1975, leis para indivíduos com algum tipo de deficiência começaram a ser feitas, em razão disso, as escolas públicas dos Estados Unidos têm sido obrigadas a incluir alunos com deficiência em classes de educação regular, onde eles possam estudar com pares com alunos sem necessidades especiais ao invés de ficaram em escolas direcionadas para os mesmos. E é claro que também existe programas para melhor capacitação de professores, para poder assim oferecer um maior suporte. Quebec/Canadá: Inicialmente é feita algumas avaliações, para jugar o nível fisiopsicológico em que o futuro aluno se encontra, partir disso a escola é responsável a iniciar uma observação da criança em sala de aula, analisando e jugando o que pode ser feito para melhor adaptação da mesma.* "

Pode ser verificado que a natureza do tipo de texto presente nesta base, é de uma natureza mais descritiva, do que os textos das bases ESSAYS e *myPersonality*, visto que os alunos irão descrever sobre atividades ou fatos de terceiros, e não sobre suas características pessoais. Esta característica representa um dos desafios desta pesquisa, que é o de verificar a viabilidade da identificação da personalidade dos alunos, baseados em textos desta natureza, ou seja, textos oriundos de atividades educacionais.

A Tabela 6.5 apresenta o formato dos dados nesta base. A coluna TEXTO apresenta os dados de todas as inserções realizadas no *Moodle* pelo aluno correspondente ao identificador ID, agrupadas em uma mesma linha. As colunas S_{Ext} a S_{Opn} apresentam um valor correspondente a cada dimensão BIG FIVE, em uma faixa de valores entre 1,00 e 5,00.

Tabela 6.5: Formato da Base UNIVERSIDADE

Id	**Nome**	**Texto**	$\mathbf{S_{Opn}}$	$\mathbf{S_{Con}}$	$\mathbf{S_{Ext}}$	$\mathbf{S_{Agr}}$	S_{Neu}
8	omitido	Partindo dos pressupostos afirmado...	3,50	4,13	3,14	3,89	4,50
9	omitido	Colocar todos os alunos na mesma s...	1,88	4,00	2,29	4,44	3,10
10	omitido	Embora as escolas brasileiras este...	1,88	4,00	2,57	4,00	3,40
11	omitido	O tema abordado é bastante comple...	2,88	3,38	2,86	2,89	4,09
13	omitido	Apesar da ideia de inclusão ser a...	2,38	2,38	3,14	3,00	3,40
15	omitido	A inclusão é vista de uma maneir...	4,50	3,38	3,43	2,78	4,20
16	omitido	"É significativo ressaltar que a ...	2,75	3,50	2,43	2,11	4,20
18	omitido	A questão abordada pelo Felipe so...	2,25	2,75	2,43	2,56	2,60
21	omitido	A inclusão é um assunto muito de...	2,13	2,88	2,43	3,56	3,20
22	omitido	Ao tentarmos apenas colocar alunos...	2,88	4,75	2,14	3,00	4,20
:	:	:	:	:	:	:	:
113	omitido	Todos os seres vivos aqui empre...	4,50	2,62	3,28	3,00	3,90

A Tabela 6.6 ilustra as informações estatísticas obtidas a partir da base UNIVERSIDADE. Inicialmente merece destaque a quantidade de registros, ou seja, o número de voluntários

Tabela 6.6: Características da Base UNIVERSIDADE

Característica	**Valor**
Idioma	Português
Linhas no Arquivo	45
Total de Usuários	45
Total de Palavras	25.066
Quantidade Média de Palavras por Usuário	557
Palavras Diferentes	4.389
Quantidade Média de Ocorrências por Palavra	5
Quantidade de Sentenças	798
Quantidade Média de Palavras por Sentença	31

representados nesta base, que é substancialmente inferior do que as bases anteriormente descritas neste capítulo. Tendo isto em consideração, esta base somente foi utilizada como base de teste, pois a realização de treinamento com um número reduzido de registros não induz a resultados aceitáveis. Esta característica remete também para outro desafio desta pesquisa, que é a utilização de bases de treinamento consolidadas, mesmo em idioma inglês, para a identificação da personalidade de uma base reduzida, no idioma português. Foi verificado que a quantidade média de palavras por sentença desta base, que é de 31, ser bem superior aos valores de 15 palavras, como no caso da base ESSAYS e 10 palavras, para a base *myPersonality*. Apesar da diferença dos idiomas, esta quantidade está associada à natureza dos textos educacionais da base UNIVERSIDADE em contrapartida aos textos mais pessoais registrados nas outras bases citadas. Também pode ser observada a quantidade bem menor de palavras diferentes encontradas nesta base, cerca de 10% das encontradas na base ESSAYS.

Na Figura 6.4 pode ser observada a distribuição das classes nesta base. Pode ser

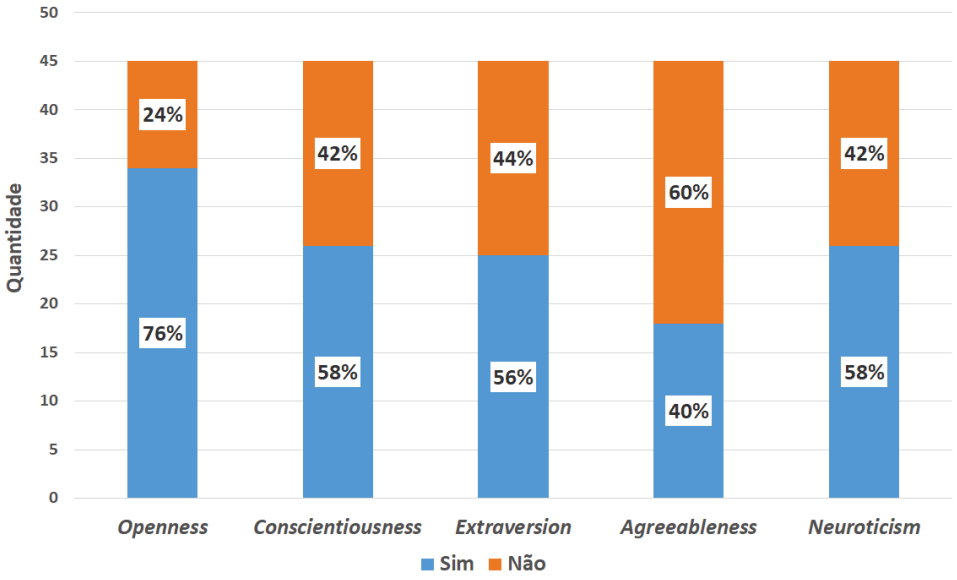


Figura 6.4: Distribuição de Classes na Base UNIVERSIDADE

verificado que a dimensão *Openness* apresenta uma grande concentração de alunos com este

perfil, o que deve ser levado em consideração na análise dos resultados obtidos na identificação desta dimensão, em face deste desbalanceamento.

6.5 Verificação da Representação com Léxico LIWC

Este experimento tem como objetivo verificar o comportamento das bases de treinamento, utilizando os classificadores isoladamente. Como forma de representação estruturada do texto, foi utilizada somente a obtenção das características a partir do Léxico LIWC, utilizando o dicionário para o idioma inglês na versão 2007. Com este processo foram obtidos vetores com 79 características.

Os testes foram realizados com os diversos classificadores selecionados, utilizando a técnica de validação cruzada *k-fold cross validation*, descrita na Seção 2.7.3, para separação das bases de treinamento e teste, com o critério de $k = 5$ e com distribuição balanceadas das classes em cada subconjunto obtido.

A Tabela 6.7 apresenta os resultados obtidos utilizando a base de dados ESSAYS. O valor da acurácia apresentada refere-se à média obtida nos processos para cada um dos classificadores, referente a cada uma das dimensões do BIG FIVE. Ao lado de cada valor de acurácia é apresentado o desvio padrão correspondente a cada série de testes. Os valores apresentados em negrito indicam os melhores resultados obtidos em cada uma das dimensões.

Tabela 6.7: Base ESSAYS com LIWC

	Openness $\overline{acc}$ (%)	***Conscientiousness*** $\overline{acc}$ (%)	***Extraversion*** $\overline{acc}$ (%)	***Agreeableness*** $\overline{acc}$ (%)	***Neuroticism*** $\overline{acc}$ (%)
BASELINE	51,6±1,7	47,1±0,9	50,6±1,2	50,3±1,5	48,6±1,2
kNN	52,9±1,6	49,4±1,4	51,0±1,5	50,4±1,7	51,1±1,5
GNB	58,8±1,4	**55,3**±1,7	52,9±1,6	**54,8**±1,4	**55,9**±1,8
LR	**59,6**±1,9	54,8±2,5	**55,0**±0,9	52,7±0,9	55,9±1,3
RFOREST	56,5±1,5	53,1±1,8	54,4±1,8	54,0±1,0	54,1±1,2
MLPC	51,5±0,1	51,8±1,6	52,3±1,1	53,0±0,1	50,1±0,5
SVM	51,7±0,5	50,5±1,7	51,3±0,4	53,1±0,1	51,8±1,9

Os valores obtidos estão coerentes com os observados na literatura investigada, apresentados na Tabela 4.1, do Capítulo 4. Pode ser observada que a maior acurácia foi obtida pelos classificadores *Gaussian Naïve Bayes* e *Logistic Regression*. Os valores obtidos para as cinco dimensões também estão próximos, indicando uma boa distribuição das classes nesta base.

A Tabela 6.8 apresenta os resultados obtidos utilizando a base de dados *myPersonality*. Neste caso pode ser observado que os melhores resultados foram obtidos nas dimensões *Openness* e *Neuroticism*, dimensões estas que apresentam um maior desbalanceamento, conforme ilustrado na Figura 6.3. Com esta base, as melhores acurácias foram obtidas com os classificadores *Logistic Regression*, *Random Forest*, *Multi-layer Perceptron* e *Support Vector Machines*.

6.6 Verificação da Representação nGRAM

Nesta seção estão apresentados os resultados obtidos com os testes de classificação realizados nas bases de treinamento, com a forma de representação nGRAM. No experimento descrito nesta seção, as bases de dados ESSAYS e *myPersonality* foram submetidas a um processo

Tabela 6.8: Base *myPersonality* com LIWC

	Openness $\overline{acc}$ (%)	***Conscientiousness*** $\overline{acc}$ (%)	***Extraversion*** $\overline{acc}$ (%)	***Agreeableness*** $\overline{acc}$ (%)	***Neuroticism*** $\overline{acc}$ (%)
BASELINE	49,9±1,4	50,5±0,9	50,3±0,7	50,0±0,5	49,2±1,0
kNN	67,1±0,7	53,5±1,0	54,0±1,1	52,2±1,0	56,4±1,4
GNB	46,6±2,3	54,2±1,0	54,2±0,9	53,0±1,3	58,6±0,9
LR	74,3±0,1	**55,2**±1,3	**58,1**±1,1	**53,9**±0,9	**62,8**±0,5
RFOREST	**74,3**±0,0	54,6±0,6	57,6±0,1	53,8±0,8	62,5±0,0
MLPC	74,3±0,0	54,1±0,0	57,2±0,7	53,1±0,0	62,5±0,0
SVM	**74,3**±0,0	54,0±0,2	57,5±0,1	53,1±0,0	62,5±0,0

de obtenção das categorias sintáticas das palavras, e geração de bases para classificação, por meio da utilização do módulo "Extrator nGRAM", conforme descrito na Seção 5.6. A verificação da acurácia dos classificadores foi realizada utilizando a técnica de validação cruzada k-fold cross validation, descrita na Seção 2.7.3, para separação das bases de treinamento e teste, com o critério de $k = 5$ e com distribuição balanceadas das classes em cada subconjunto obtido. Os valores apresentados representam a acurácia média obtida em cada configuração.

6.6.1 Base ESSAYS

Com a aplicação desta técnica na base de dados ESSAYS, foram obtidas 10 características, utilizando *unigram*, 110 características aplicando *bigram* e 1103 características com *trigram*.

6.6.1.1 Openness

A Figura 6.5 apresenta os resultados obtidos com a classificação realizada com representação do texto em *unigram*, *bigram* e *trigram*, na dimensão *Openness*.

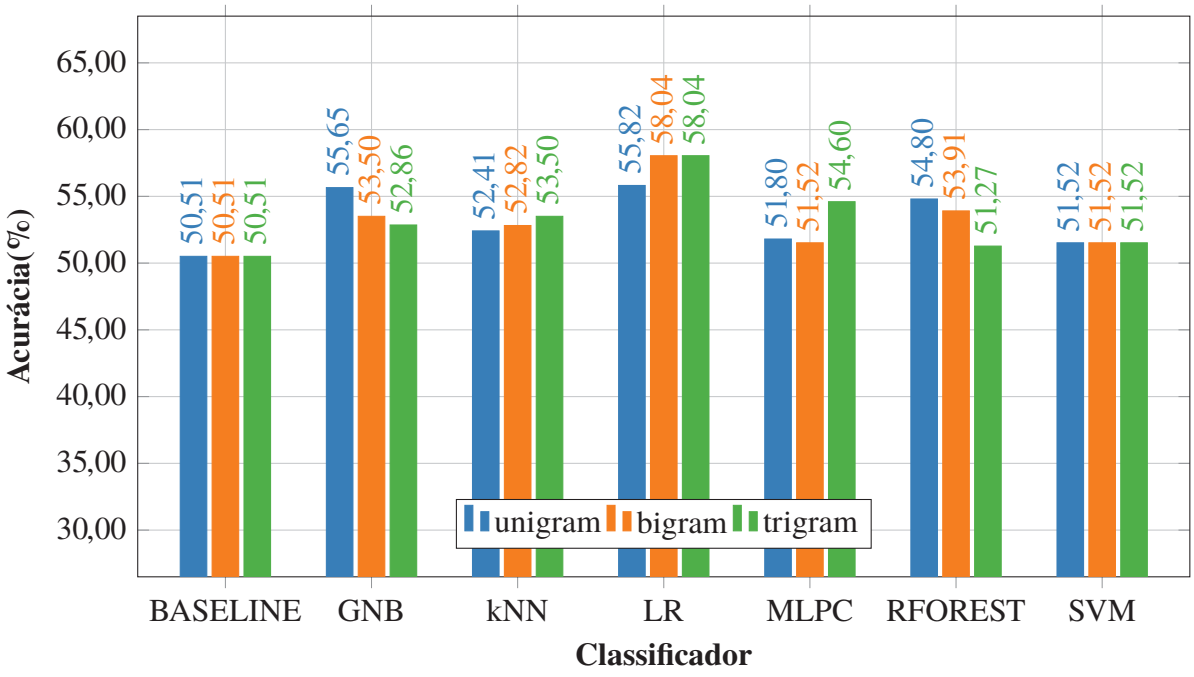


Figura 6.5: Comparativo *Openness* da Base ESSAYS com nGRAM

Estes resultados demonstram que não ocorreu uma variação significativa entre os três métodos. Os resultados da acurácia observada com a aplicação do modelo de representação nGRAM, para esta dimensão de personalidade, indicam resultados inferiores do que com a representação utilizando o léxico LIWC, conforme observado na Tabela 6.7.

O melhor desempenho foi obtido com o classificador *Logistic Regression* seguido do *Gaussian Naïve Bayes*. No caso dos classificadores *Gaussian Naïve Bayes* e *Random Forest*, foi ainda verificado que a forma de representação *unigram* teve resultados melhores do que as outras duas formas que utilizam mais características.

6.6.1.2 Conscientiousness

Os resultados do processo de classificação da dimensão *Conscientiousness*, estão apresentados na Figura 6.6. Nesta dimensão foi observado novamente que o melhor desempenho foi obtido com o classificador *Logistic Regression*, seguido do *Gaussian Naïve Bayes*. Mesmo assim os resultados obtidos ficaram inferiores em relação aos obtidos com a representação como o léxico LIWC. Também não foi observada variação significativa com a utilização dos três tipos de nGRAM.

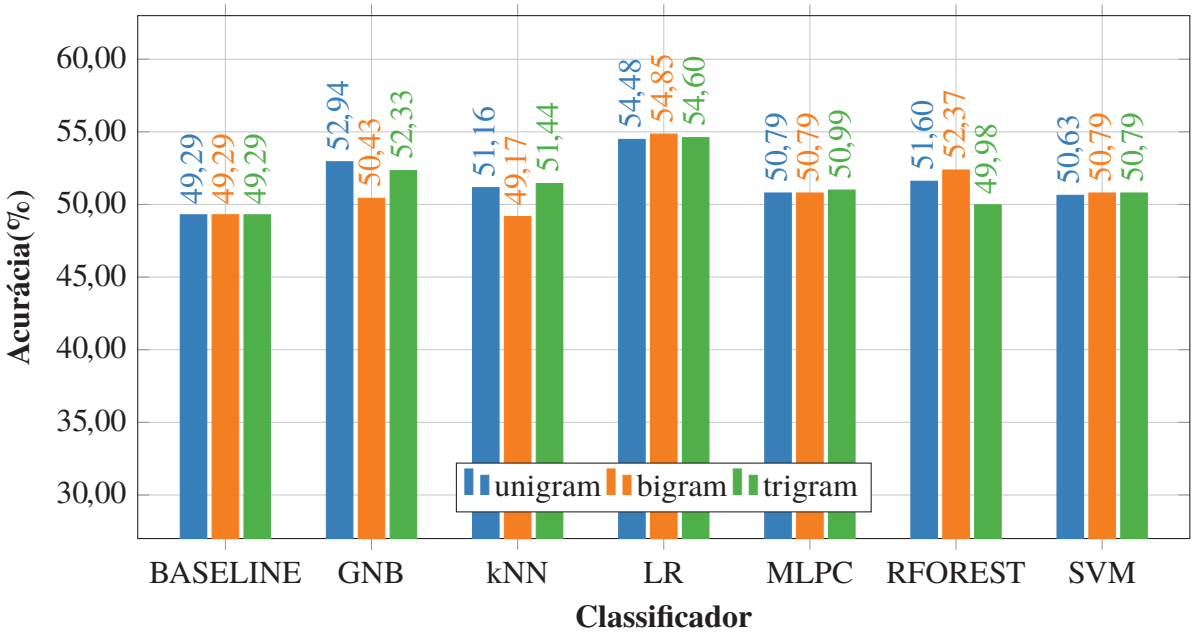


Figura 6.6: Comparativo *Conscientiousness* da Base ESSAYS com nGRAM

6.6.1.3 *Extraversion*

A Figura 6.7 ilustra os resultados obtidos com a dimensão *Extraversion*. Neste caso foi verificado que o classificador *Logistic Regression* apresentou melhores resultados, e ligeiramente superiores aos obtidos com o léxico LIWC.

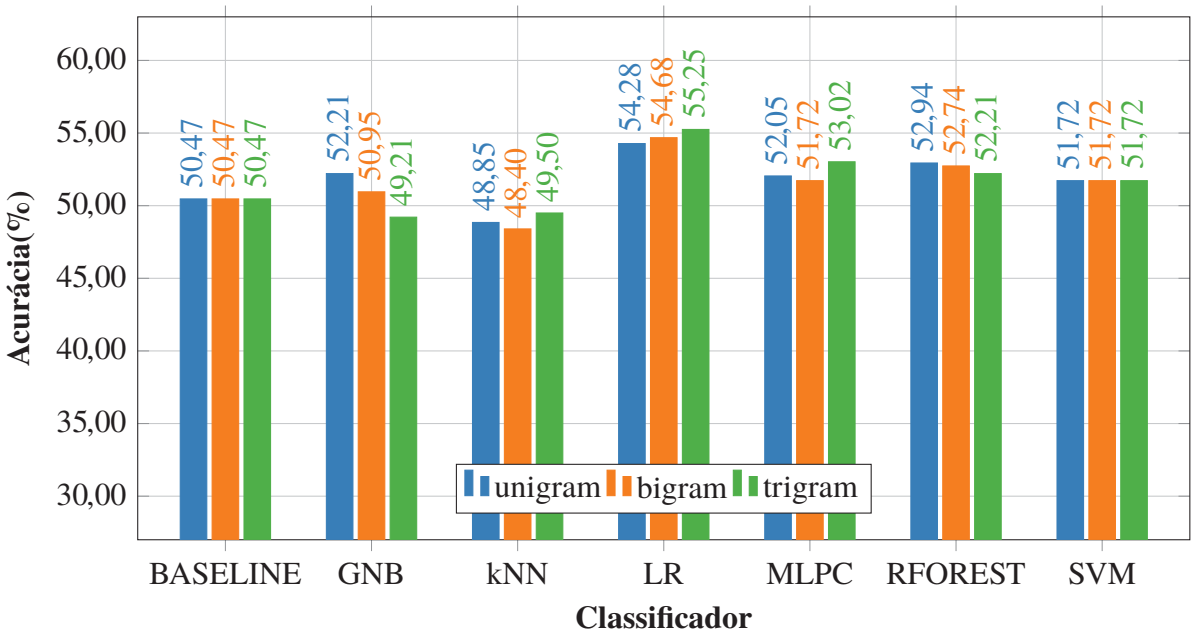


Figura 6.7: Comparativo *Extraversion* da Base ESSAYS com nGRAM

6.6.1.4 *Agreeableness*

A Figura 6.8 ilustra os resultados obtidos com a dimensão *Agreeableness*.

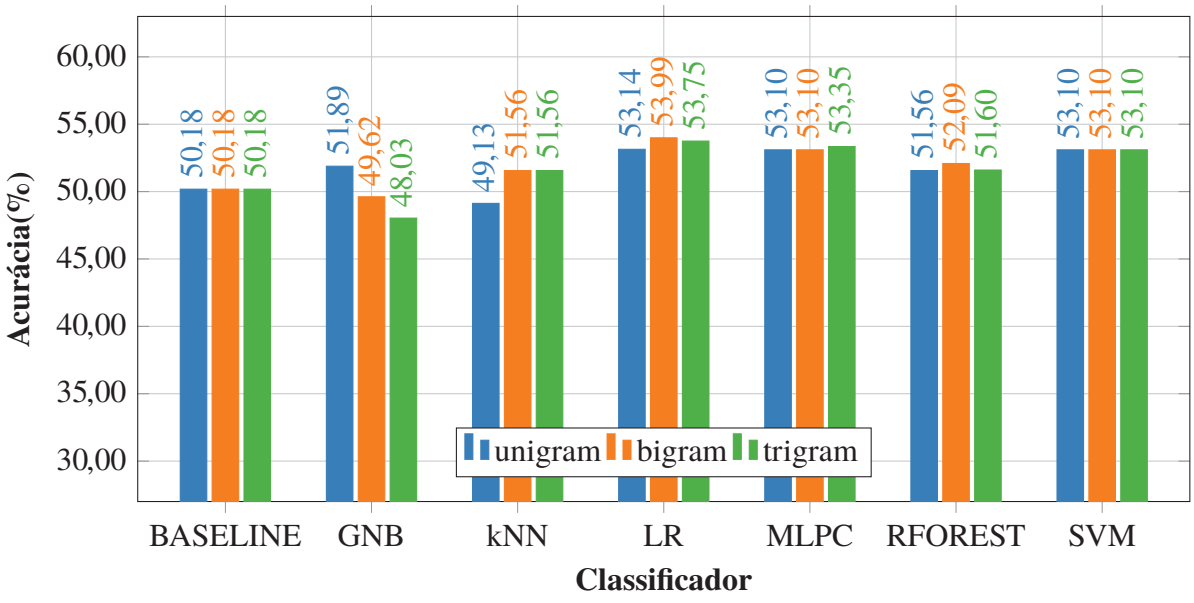


Figura 6.8: Comparativo *Agreeableness* da Base ESSAYS com nGRAM

No caso desta dimensão o comportamento dos classificadores ficou próximo para os três tipos de nGRAM, exceto para o classificador *Gaussian Naïve Bayes*, que apresentou uma relação decrescente de acurácia.

6.6.1.5 Neuroticism

A Figura 6.9 ilustra os resultados obtidos com a dimensão *Neuroticism*. No caso do *Neuroticism*, o classificador *Multi-layer Perceptron* apresentou acurácia ligeiramente crescente com o aumento do grau do nGRAM, sendo que para os demais ocorreu comportamento similar.

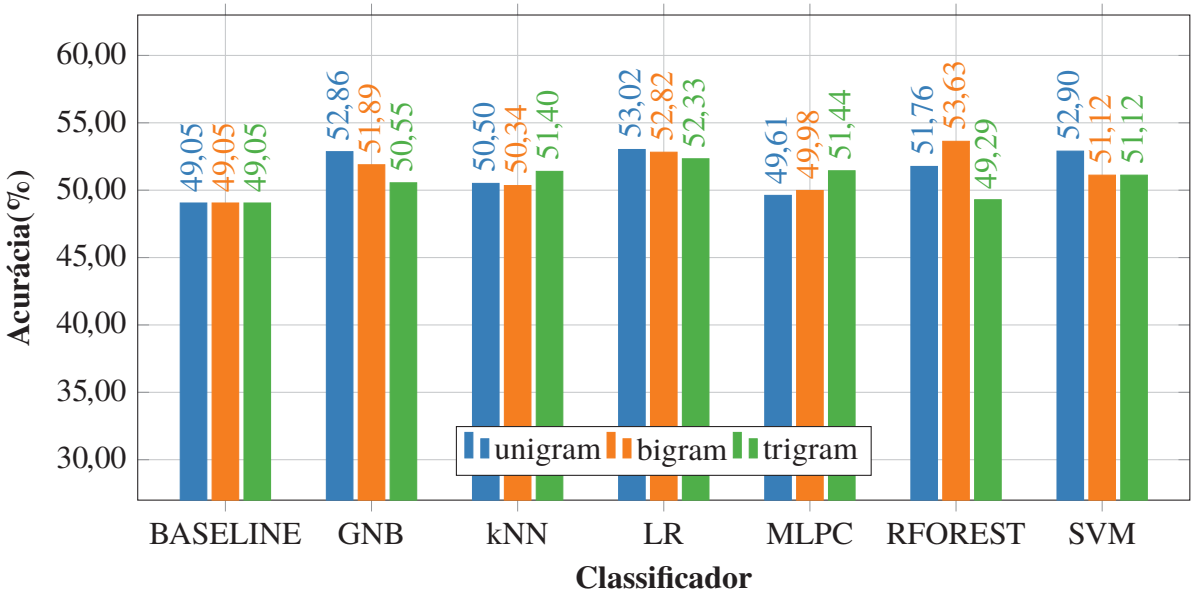


Figura 6.9: Comparativo *Neuroticism* da Base ESSAYS com nGRAM

6.6.2 Base *myPersonality*

No caso da base de dados *myPersonality*, foram obtidas 10 características, utilizando *unigram*, 110 características aplicando *bigram* e 1025 características com *trigram*.

6.6.2.1 *Openness*

A Figura 6.10 ilustra os resultados obtidos na dimensão *Openness*. Nesta base, os resultados observados para a acurácia na dimensão *Openness* foram substancialmente superiores do que os obtidos na base ESSAYS, com os classificadores *Logistic Regression*, *Multi-layer Perceptron* e *Support Vector Machines*. No caso do classificador *Gaussian Naïve Bayes* foram verificados resultados muito baixos, quando da aplicação de *unigram* e *bigram*.

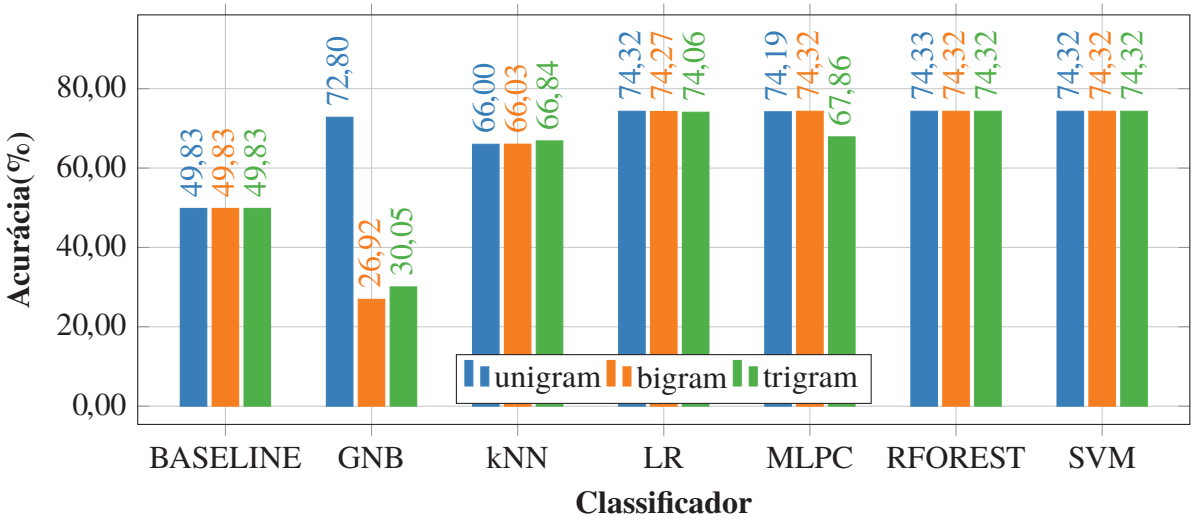


Figura 6.10: Comparativo *Openness* da Base *myPersonality* com nGRAM

6.6.2.2 *Conscientiousness*

A Figura 6.11 apresenta os resultados obtidos com o experimento aplicado à dimensão *Conscientiousness*. Para este caso, os resultados verificados para a acurácia, ficaram em torno de 50% para todos os classificadores, não ocorrendo relevante variação em relação ao tipo de nGRAM utilizado.

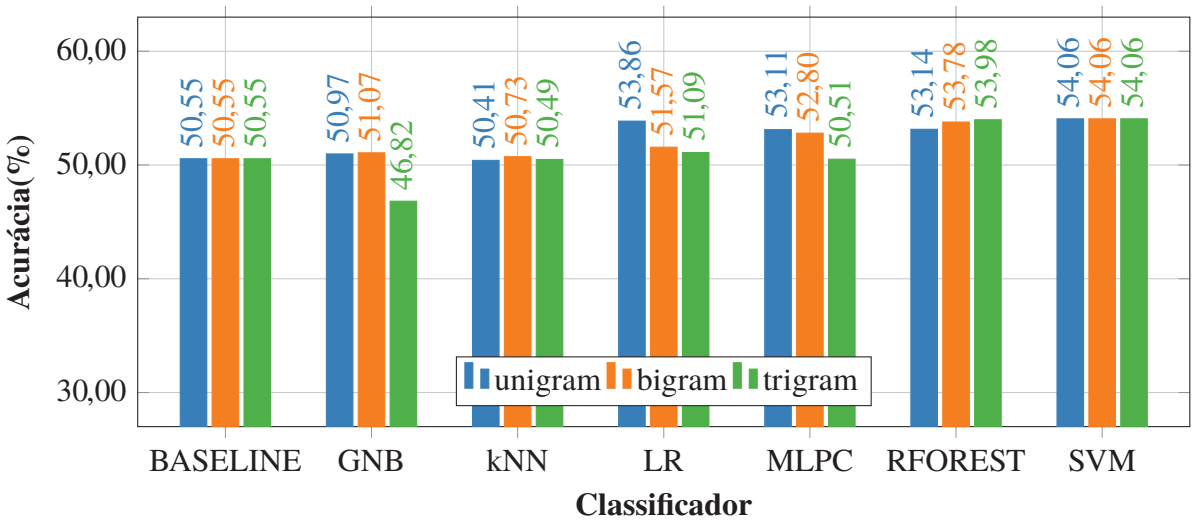


Figura 6.11: Comparativo *Conscientiousness* da Base *myPersonality* com nGRAM

6.6.2.3 *Extraversion*

A Figura 6.12 ilustra os resultados obtidos com a dimensão *Extraversion*. Para o

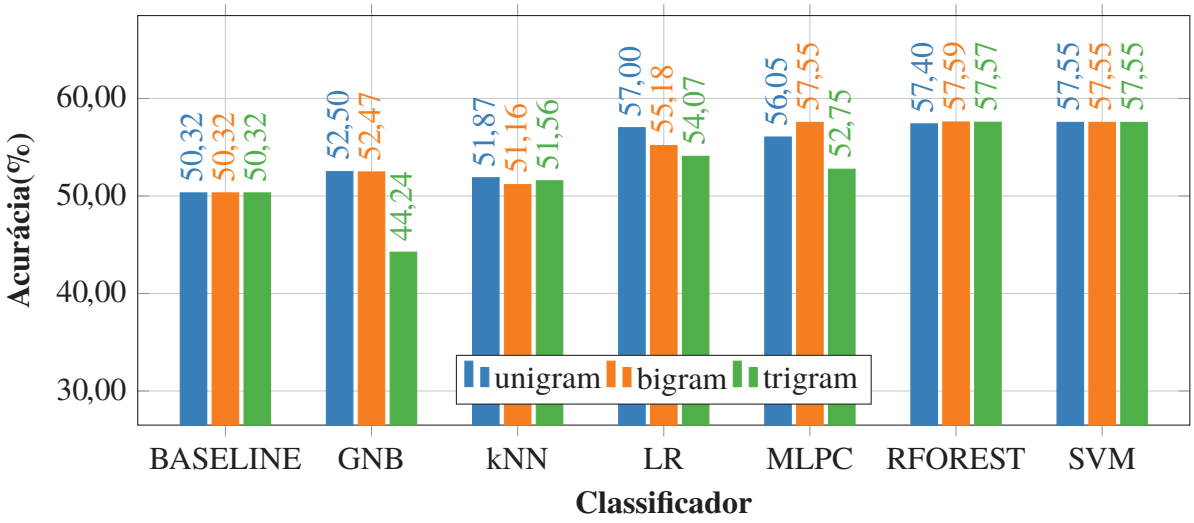


Figura 6.12: Comparativo *Extraversion* da Base *myPersonality* com nGRAM

caso da *Extraversion*, houve decréscimo da acurácia, no caso dos classificadores *Gaussian Naïve Bayes*, *Logistic Regression* e *Multi-layer Perceptron*, sendo praticamente iguais nos demais classificadores, em relação ao grau de *n-gram*. O classificador *Support Vector Machines* apresentou uma acurácia mais significativa, na faixa de 57%, para os três casos de nGRAM, valores estes também observados no caso do *Logistic Regression* com *unigram* e também do *Multi-layer Perceptron* com *bigram*.

6.6.2.4 *Agreeableness*

A Figura 6.13 ilustra os resultados obtidos com a dimensão *Agreeableness*. Neste caso, também foi verificada uma acurácia em torno de 50%, sendo ligeiramente mais expressiva no caso dos classificadores *Multi-layer Perceptron* e *Support Vector Machines*.

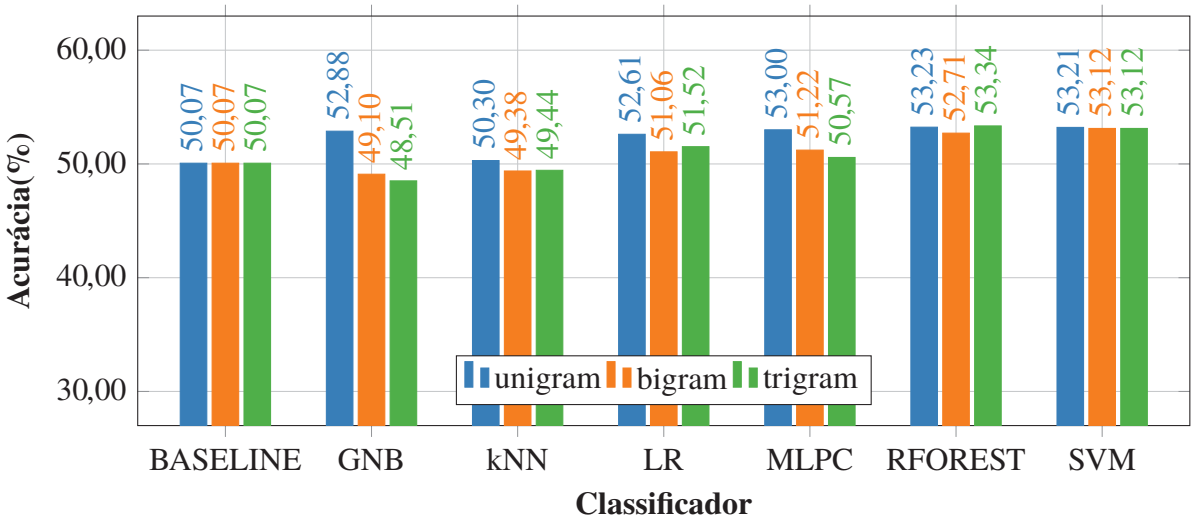


Figura 6.13: Comparativo *Agreeableness* da Base *myPersonality* com nGRAM

6.6.2.5 *Neuroticism*

A Figura 6.14 apresenta os resultados obtidos com a dimensão *Neuroticism*. Para o *Neuroticism* foi verificada uma acurácia mais acentuada, acima de 60%, nos classificadores *Logistic Regression*, *Multi-layer Perceptron* e *Support Vector Machines*. Esta condição também foi verificada no caso do *Gaussian Naïve Bayes*, mas somente para o *unigram*, sendo que para *bigram* e *trigram* os valores foram muito baixo em relação aos demais classificadores.

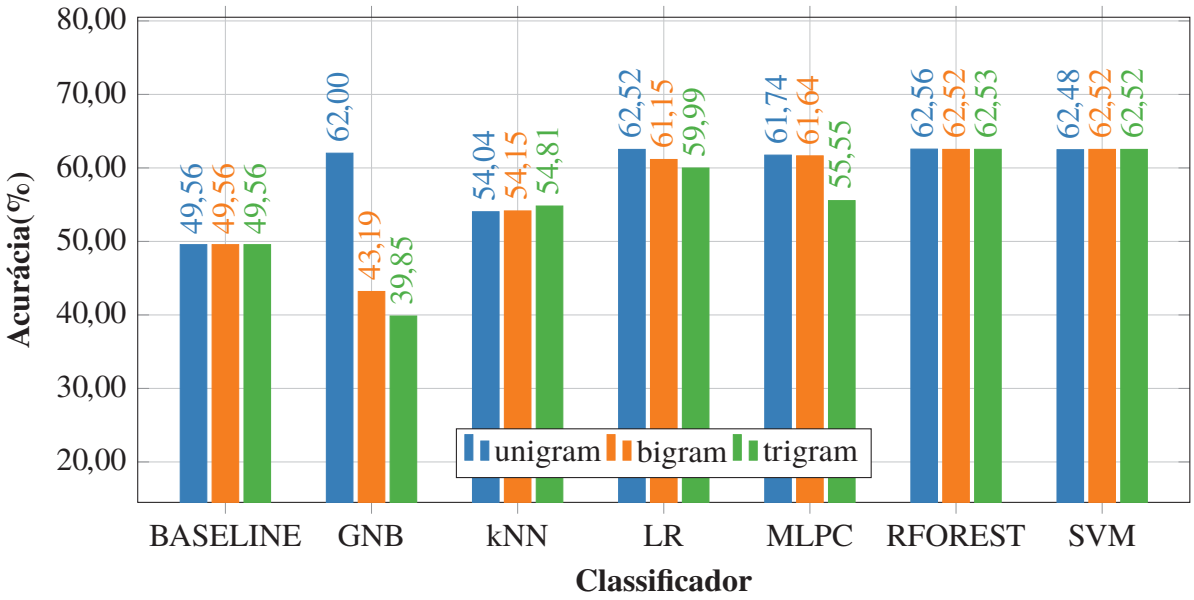


Figura 6.14: Comparativo *Neuroticism* da Base *myPersonality* com nGRAM

6.6.3 Considerações

Em termos gerais, a utilização da estruturação do texto com técnicas do nGRAM, não demonstrou resultados mais promissores do que os resultados obtidos com o léxico LIWC. Estes resultados estão coerentes com outros experimentos (Majumder et al., 2017), que também verificaram que a utilização desta forma de representação, de forma isolada, não supera a representação do texto utilizando o léxico.

O experimento realizado na presente pesquisa procurou verificar quais o resultados obtidos com a representação do texto utilizando nGRAM, aplicados sobre as categorias sintáticas identificadas no módulo "Categorizador", conforme apresentado na Seção 5.5 e proposto no modelo IP3. Foi da mesma forma, constatado que a aplicação isolada desta forma de representação, não teria melhores resultados do que somente a utilização do LIWC. Mas como foi verificado no experimento apresentado na Seção 6.8, esta forma de representação utilizada em conjunto com o LIWC, acarreta ganhos em diversos cenários.

6.7 Verificação da Representação *Word2Vec*

Nesta seção estão apresentados os resultados obtidos com os testes de classificação realizados nas bases de treinamento, com a forma de representação *Word2Vec*. No experimento descrito nesta seção, as bases de dados ESSAYS e *myPersonality* foram submetidas a um processo de obtenção das categorias sintáticas das palavras, utilizado o módulo "Categorizador" e geração

de bases para classificação, por meio da utilização do módulo "Extrator *Word2Vec*", conforme descrito na Seção 5.7. O objetivo deste experimento é verificar a aplicação da técnica de *Word2Vec* como forma de representação do texto em processo de identificação de personalidade utilizando técnicas de aprendizado de máquina, bem como avaliar a influência da quantidade de características geradas nos resultados obtidos.

As Seções 6.7.1 e 6.7.2 apresentam os resultados obtidos, na forma de um gráfico para cada dimensão de personalidade, para as duas bases avaliadas. Foi utilizado o conjunto de classificadores definido para o experimento, sendo gerados vetores de representação nos tamanhos 10, 20, 50, 100, 200, 300, 400 e 500. A verificação da acurácia dos classificadores foi realizada utilizando a técnica de validação cruzada k-fold cross validation, descrita na Seção 2.7.3, para separação das bases de treinamento e teste, com o critério de $k = 5$ e com distribuição balanceada das classes em cada subconjunto obtido. Os valores apresentados representam a acurácia média obtida em cada configuração. A identificação do tamanho mínimo de características a ser utilizado irá influenciar diretamente no tempo de processamento a ser utilizado pelos classificadores, sendo que um aumento deste custo computacional sem uma melhoria razoável na acurácia, não seria justificável.

6.7.1 Base ESSAYS

Nesta seção são apresentados os resultados do experimento realizado com a base ESSAYS.

6.7.1.1 Openness

A Figura 6.15 ilustra os valores encontrados com a dimensão *Openness*.

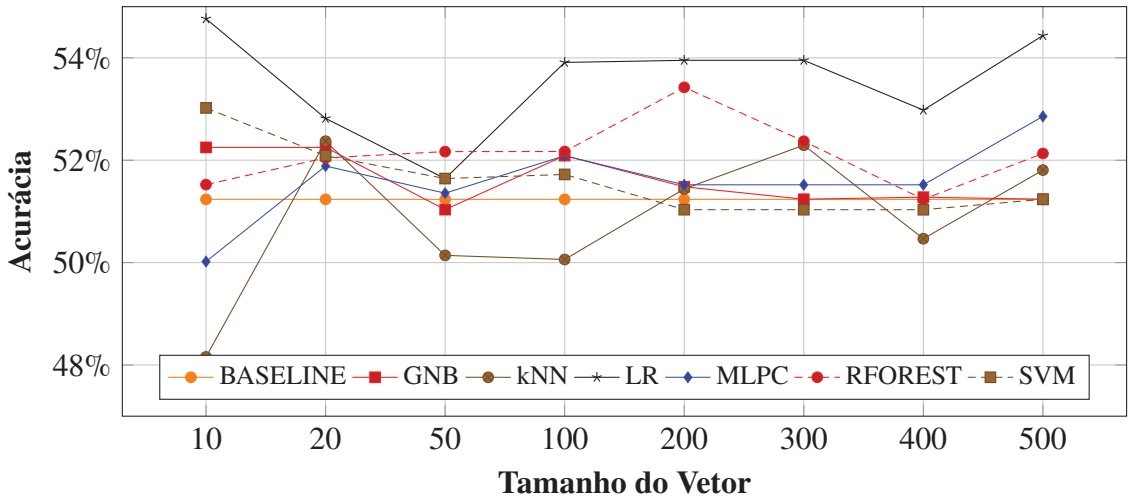


Figura 6.15: Comparativo *Openness* da Base ESSAYS com *Word2Vec*

Para esta dimensão, o classificador *Logistic Regression* apresentou a melhor acurácia, sendo que os demais apresentaram um pouco abaixo. Foi verificada pouca variação nos resultados em função do tamanho do vetor, principalmente a partir do tamanho 200.

6.7.1.2 *Conscientiousness*

Na Figura 6.16 estão apresentados os resultados da dimensão *Conscientiousness*. Nesta situação, não foi verificado um destaque para um classificador específico, sendo que o *Gaussian Naïve Bayes*, o *Random Forest*, o *Logistic Regression* e o *Multi-layer Perceptron* tiveram resultados inferiores ao *Baseline* em diversas situações.

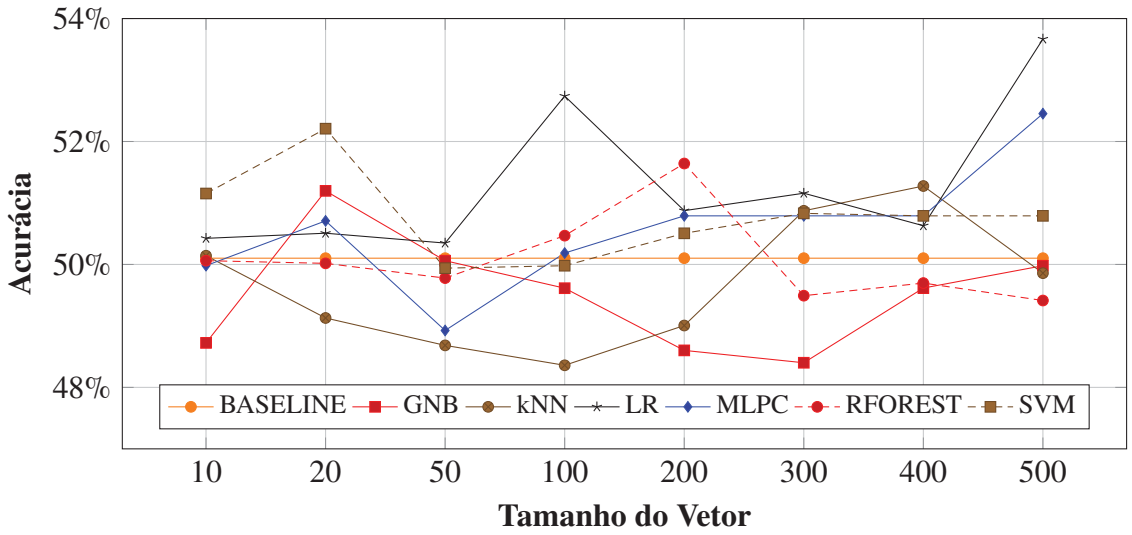


Figura 6.16: Comparativo *Conscientiousness* da Base ESSAYS com *Word2Vec*

6.7.1.3 *Extraversion*

A Figura 6.17 ilustra os valores encontrados com a dimensão *Extraversion*. Nesta condição os resultados ficaram em torno e um pouco abaixo do *Baseline*, não demonstrando ganhos nesta configuração.

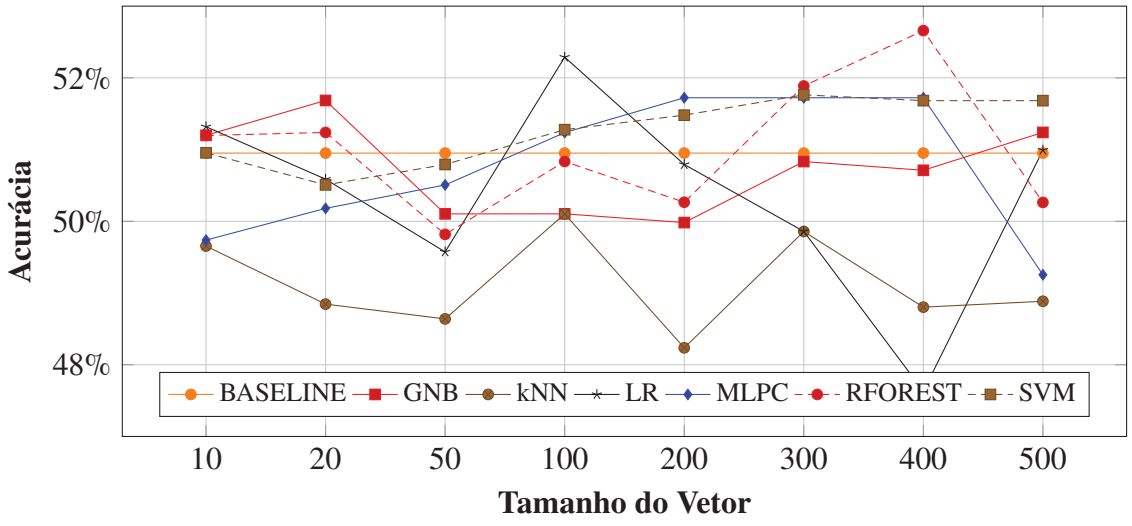


Figura 6.17: Comparativo *Extraversion* da Base ESSAYS com *Word2Vec*

6.7.1.4 *Agreeableness*

A figura 6.18 apresenta o comportamento para a dimensão *Agreeableness*. O classificador *Support Vector Machines* apresentou um comportamento regular, em torno de 53% de acurácia, assim como o *Multi-layer Perceptron* com tamanhos entre 200 e 400. Já no caso do *Gaussian Naïve Bayes* e *k-nearest neighbors* os resultados ficaram praticamente abaixo do *Baseline*.

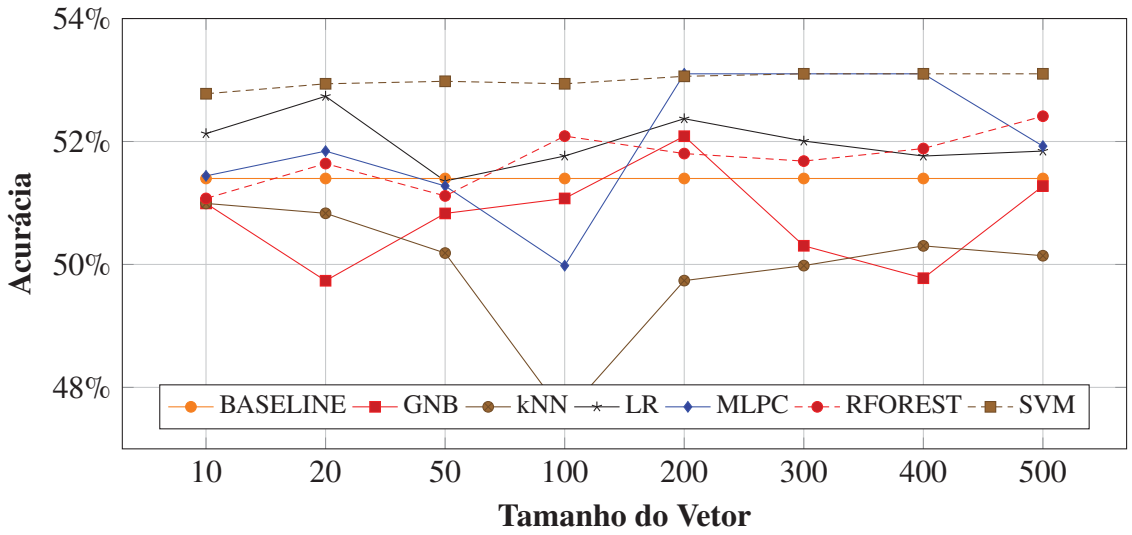


Figura 6.18: Comparativo *Agreeableness* da Base ESSAYS com *Word2Vec*

6.7.1.5 *Neuroticism*

A Figura 6.19 ilustra os valores encontrados com a dimensão *Neuroticism*. Esta foi a única dimensão em que os resultados ficaram acima do *Baseline* praticamente em todos os casos, sendo que o melhor desempenho foi observado no caso do classificador *Logistic Regression*.

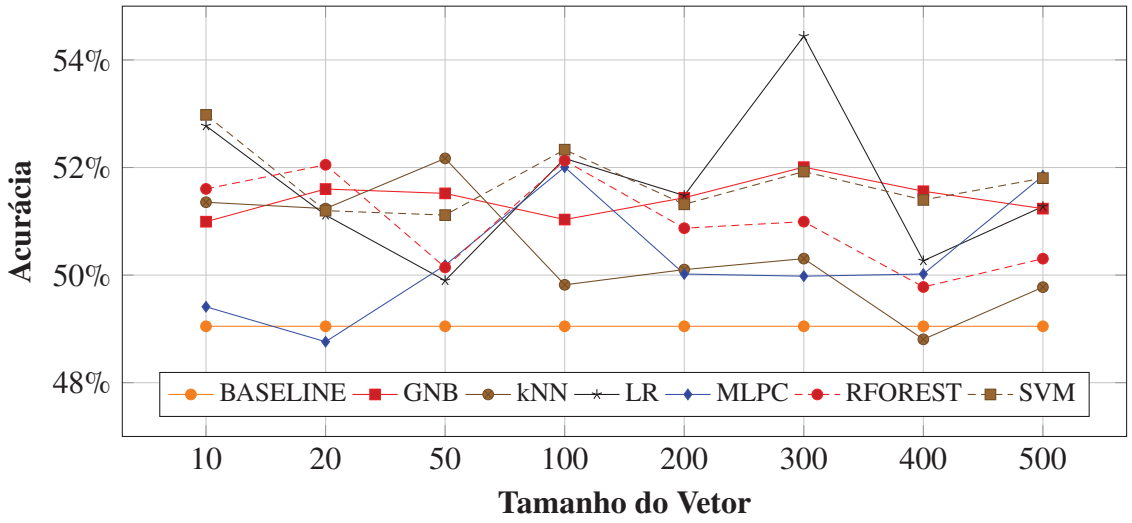


Figura 6.19: Comparativo *Neuroticism* da Base ESSAYS com *Word2Vec*

6.7.2 Base *myPersonality*

Nesta seção são apresentados os resultados do experimento realizado com a base *myPersonality*.

6.7.2.1 Openness

A Figura 6.20 ilustra os valores encontrados com a dimensão *Openness*. Nesta dimensão foi observado um comportamento linear, com resultados próximos aos obtidos com o experimento realizado com o LIWC. No caso do classificador *Gaussian Naïve Bayes* houve uma relação inversa entre a acurácia e o tamanho do vetor de características.

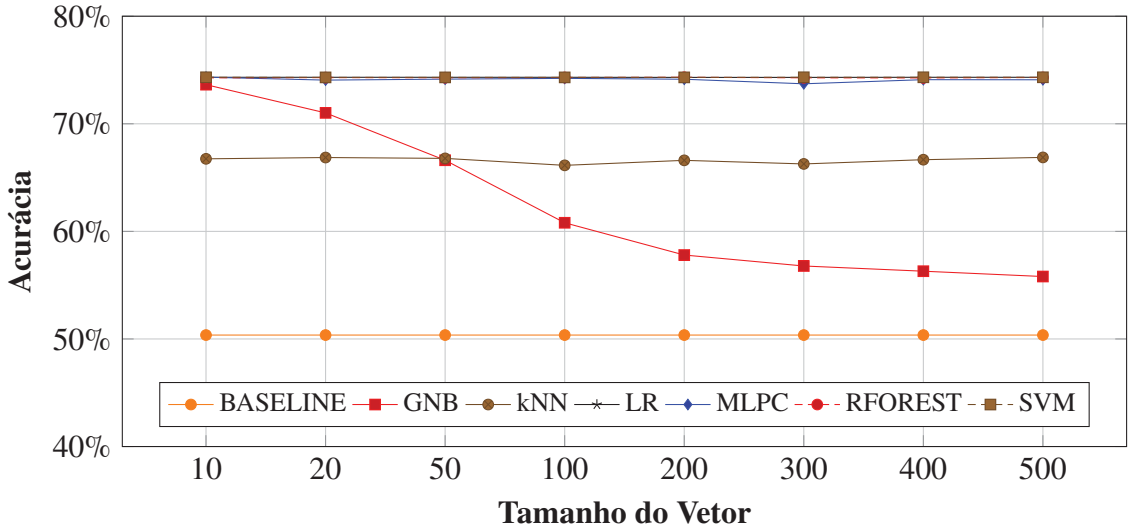


Figura 6.20: Comparativo *Openness* da Base *myPersonality* com *Word2Vec*

6.7.2.2 Conscientiousness

Na Figura 6.21 estão apresentados os resultados da dimensão *Conscientiousness*. Para

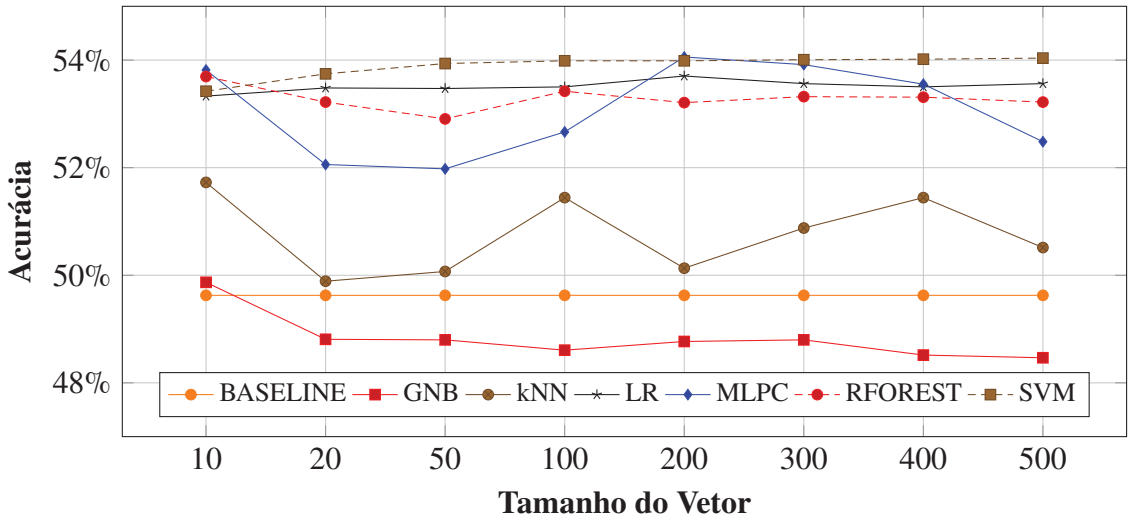


Figura 6.21: Comparativo *Conscientiousness* da Base *myPersonality* com *Word2Vec*

este caso, os classificadores *Random Forest* e *Support Vector Machines* apresentaram os melhores resultados, sem grandes variações com vetor maior que 200, e apresentando resultados próximos aos obtidos como LIWC para estes classificadores.

6.7.2.3 Extraversion

A Figura 6.22 ilustra os valores encontrados com a dimensão *Extraversion*. No caso da *Extraversion*, os classificadores *Multi-layer Perceptron*, Random Forest e *Support Vector Machines* obtiveram os melhores resultados, com estabilidade em relação ao tamanho do vetor, e o classificador *Gaussian Naïve Bayes* novamente demonstrou uma baixa acurácia.

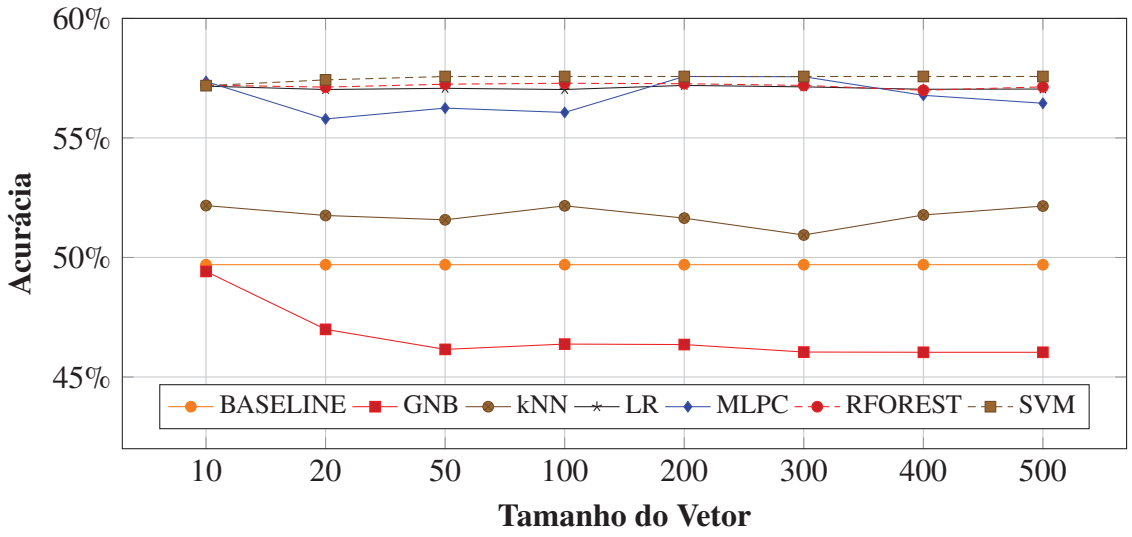


Figura 6.22: Comparativo *Extraversion* da Base *myPersonality* com *Word2Vec*

6.7.2.4 Agreeableness

A figura 6.23 apresenta o comportamento para a dimensão *Agreeableness*. Nesta

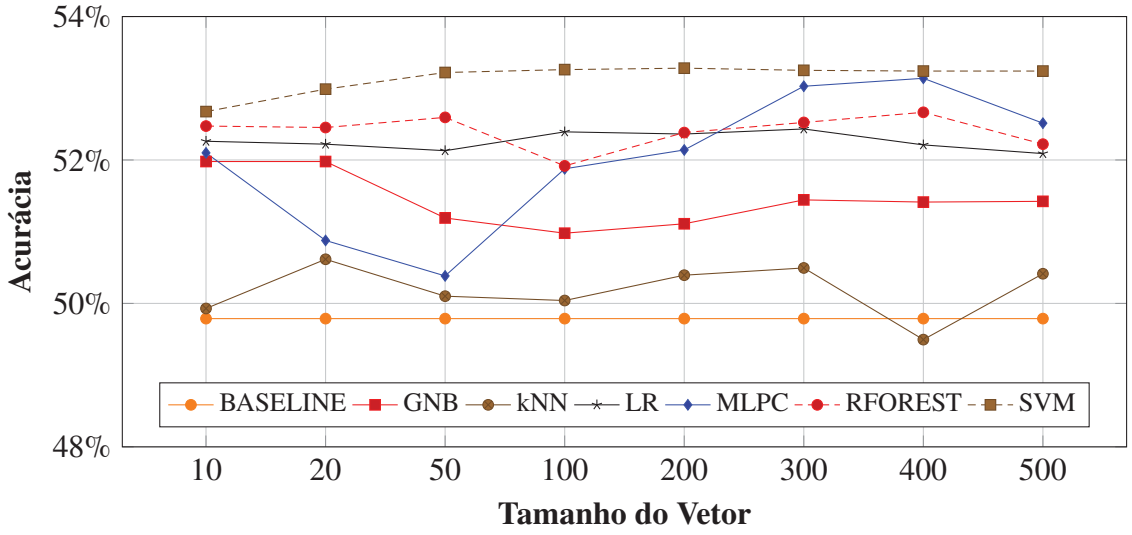


Figura 6.23: Comparativo *Agreeableness* da Base *myPersonality* com *Word2Vec*

dimensão os classificadores obtiveram resultados superiores ao *Baseline* em praticamente todas as combinações, mas mesmo assim, apresentando resultados próximos ao LIWC. O destaque para esta dimensão ficou com os classificadores *Random Forest* e *Support Vector Machines*, acompanhados pelo classificador *Multi-layer Perceptron* a partir do tamanho do vetor igual a 100.

6.7.2.5 Neuroticism

A Figura 6.24 ilustra os valores encontrados com a dimensão *Neuroticism*. Foi observado um comportamento semelhante entra a dimensão *Neuroticism* e a dimensão *Openness*, para esta base. Os classificadores *Multi-layer Perceptron*, Random Forest, *Support Vector Machines* e *Logistic Regression* apresentaram os melhores resultados, sendo estáveis em relação ao tamanho do vetor, mas ainda assim com resultados próximos aos obtidos com o LIWC. Novamente o classificador *Gaussian Naïve Bayes* apresentou resultados inferiores e decréscimo da acurácia com o aumento do tamanho do vetor de características.

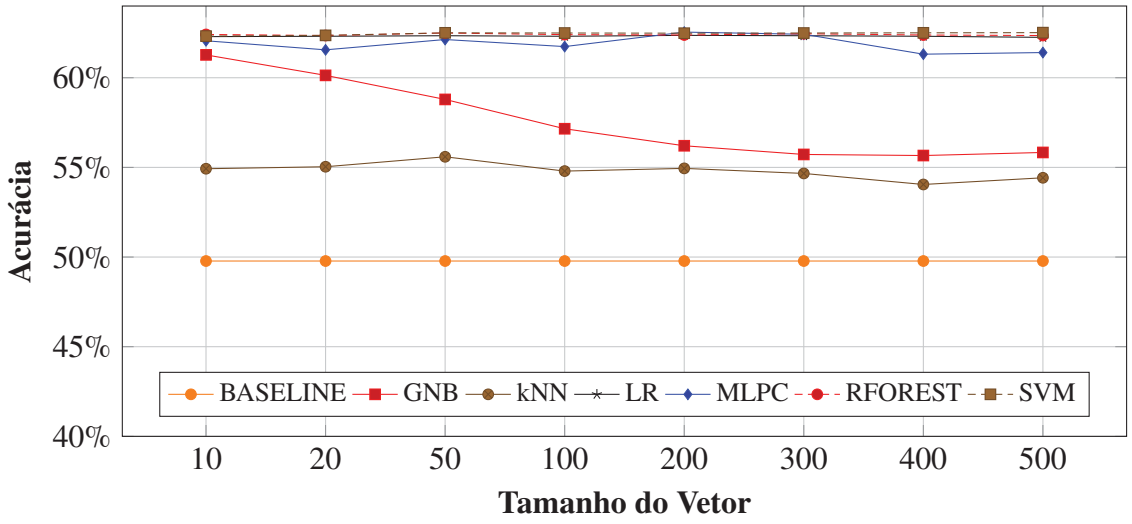


Figura 6.24: Comparativo *Neuroticism* da Base *myPersonality* com *Word2Vec*

6.7.3 Considerações

Em termos gerais, a utilização da estruturação do texto com técnicas *Word2Vec* também não demonstrou resultados mais promissores do que os resultados obtidos com o léxico LIWC. Isto ocorreu de forma mais acentuada em relação à base *myPersonality*. O experimento realizado nesta pesquisa procurou verificar quais o resultados obtidos com a representação do texto utilizando *Word2Vec*, aplicados sobre as categorias sintáticas identificadas no módulo "Categorizador", conforme apresentado na Seção 5.5, e proposto no modelo IP3. Foi da mesma forma, constatado que a aplicação isolada desta forma de representação, não teria melhores resultados do que somente a utilização do LIWC. Outro ponto observado foi de que o aumento do tamanho do vetor de características não indicou melhoria nos resultados que justificassem a utilização de vetores com tamanho elevado. Partindo desta constatação, para o experimento apresentado na Seção 6.8, foi estipulado um valor intermediário de 100 características para a representação nGRAM, sendo constatado que esta forma de representação utilizada em conjunto com o LIWC, acarreta ganhos em diversos cenários.

6.8 Validação do Modelo IP3

Nas Seções 6.5, 6.6 e 6.7 foram apresentados os resultados obtidos com a aplicação das técnicas de representação, de forma isolada, na classificação das bases ESSAYS e *myPersonality*, utilizando a separação de cada uma das bases em um conjunto de treinamento e um conjunto de teste. O experimento descrito nesta seção tem como objetivo utilizar o modelo IP3 para a realização da identificação da personalidade dos alunos, de acordo com os textos presentes na base UNIVERSIDADE, e comparar com os valores identificados no levantamento manual, conforme descrito na Seção 5.10.

Neste cenário, as bases ESSAYS e *myPersonality* são utilizadas somente para o treinamento dos classificadores, enquanto a base UNIVERSIDADE é utilizada como base de teste. Os valores inferidos pelas diversas composições de classificadores e técnicas de representação, foi comparada com os valores OCEAN obtidos a partir da identificação manual da personalidade dos alunos. Foi então verificada a acurácia obtida em cada caso, conforme apresentado nas seções seguintes deste capítulo.

Para cada uma das dimensões OCEAN, foi realizada uma série de combinações entre técnicas de representação, bases de treinamento e classificadores. Cada uma destas combinações, denominada de CONJUNTO, pode conter um ou mais agrupamentos de base de treinamento, tipo de representação e algoritmo de classificação, no modelo *Ensemble* de agrupamento de classificadores.

A acurácia apresentada para cada CONJUNTO é obtida a partir da comparação entre os valores estimados pelo CONJUNTO de classificadores em questão e os valores obtidos na identificação manual presente na base UNIVERSIDADE.

Os CONJUNTOS selecionados nesta seção apresentam as combinações que apresentam os melhores resultados encontrados nos experimentos realizados. Pode ser observado que os valores obtidos nas seções anteriores, nos experimentos que utilizaram somente as bases bases ESSAYS e *myPersonality* são diferentes dos resultados obtidos quando da inferência do perfil de personalidade dos alunos presentes na base UNIVERSIDADE.

Por exemplo, o valor da acurácia de 54,8% obtido utilizando o classificador Gaussian Naïve Bayes com a representação LIWC aplicada sobre a base ESSAYS, na dimensão *Agreeableness*, apresentado na Tabela 6.7 é diferente do valor de 60,0% obtido no CONJUNTO Agr2 indicado na Tabela 6.12. Apesar de utilizar a mesma combinação para classificação, este CONJUNTO teve com base de teste a base UNIVERSIDADE e não um subconjunto da base ESSAYS.

6.8.1 Resultados obtidos com a dimensão *Openness*

A Tabela 6.9 apresenta os resultados obtidos com 10 configurações específicas de conjuntos de classificadores, combinando as duas bases de treinamento utilizadas e diversos modos de representação do texto.

Tabela 6.9: Resultados do *Ensemble* de Classificação da Dimensão *Openness*

Conjunto	**Acurácia**	**Classificador**	**Modo**	**Base**
Opn1	46,67%	kNN	LIWC	ESSAYS
Opn2	80,00%	RFOREST	LIWC	ESSAYS
Opn3	88,89%	MLPC	LIWC	ESSAYS
Opn4	75,56%	MLPC	*unigram*	ESSAYS
Opn5	88,89%	SVM	*Word2Vec*	ESSAYS
Opn6	66,67%	RFOREST	LIWC	ESSAYS
		MLPC	LIWC	ESSAYS
Opn7	75,56%	RFOREST	LIWC	ESSAYS
		MLPC	LIWC	ESSAYS
		MLPC	*unigram*	ESSAYS
Opn8	**91,11%**	RFOREST	LIWC	ESSAYS
		MLPC	LIWC	ESSAYS
		MLPC	*unigram*	ESSAYS
		SVM	*Word2Vec*	ESSAYS
Opn9	57,78%	kNN	LIWC	*myPersonality*
		RFOREST	*unigram*	*myPersonality*
Opn10	75,56%	MLPC	LIWC	*myPersonality*
		GNB	LIWC	*myPersonality*
		LR	*unigram*	*myPersonality*

Os melhores resultados foram obtidos com a base de treinamento ESSAYS, sendo que a combinação das três formas de representação, LIWC, *unigram* e *Word2Vec*, no conjunto "Opn8", apresentou o melhor resultado, com acurácia de 91%. Estes valores obtidos superam os valores identificados na literatura investigada, em que o melhor resultado observado, foi de 74%, obtido com a representação LIWC, utilizando a base *myPersonality* e redes neurais, no trabalho apresentado por Tandera et al. (2017), conforme apresentado na Tabela 4.1.

6.8.2 Resultados obtidos com a dimensão *Conscientiousness*

A Tabela 6.10 apresenta os resultados obtidos com 13 configurações específicas de conjuntos de classificadores, combinando as duas bases de treinamento utilizadas e diversos modos de representação do texto.

Tabela 6.10: Resultados do *Ensemble* de Classificação da Dimensão *Conscientiousness*

Conjunto	**Acurácia**	**Classificador**	**Modo**	**Base**
Con1	51,11%	RFOREST	LIWC	ESSAYS
Con2	48,89%	SVM	LIWC	ESSAYS
Con3	57,78%	SVM	*unigram*	ESSAYS
Con4	53,33%	kNN	*Word2Vec*	ESSAYS
Con5	55,56%	kNN	*unigram*	ESSAYS
		LR	*unigram*	ESSAYS
Con6	53,33%	RFOREST	LIWC	ESSAYS
		kNN	LIWC	ESSAYS
Con7	57,78%	RFOREST	LIWC	ESSAYS
		kNN	LIWC	ESSAYS
		MLPC	*unigram*	ESSAYS
Con8	**62,22%**	RFOREST	LIWC	ESSAYS
		kNN	LIWC	ESSAYS
		MLPC	*unigram*	ESSAYS
		SVM	*Word2Vec*	ESSAYS
Con9	48,89%	kNN	LIWC	*myPersonality*
Con10	42,22%	MLPC	LIWC	*myPersonality*
Con11	53,33%	GNB	*Word2Vec*	*myPersonality*
Con12	42,22%	GNB	*unigram*	*myPersonality*
Con13	42,22%	GNB	*unigram*	*myPersonality*
		MLPC	*unigram*	*myPersonality*

Para esta dimensão, novamente os melhores resultados foram obtidos com a utilização da base ESSAYS para treinamento dos classificadores, bem como o melhor resultado foi obtido com a combinação das três técnicas de representação. O melhor resultado observado foi com o conjunto “Con8”, que apresentou uma acurácia de 62%.

Na literatura investigada, o melhor resultado encontrado foi no trabalho de Alam et al. (2013), que utilizando o classificador *Multinomial Naïve Bayes*, na base *myPersonality*, com representação nGRAM, chegou a uma acurácia de 59%.

6.8.3 Resultados obtidos com a dimensão *Extraversion*

A Tabela 6.11 apresenta os resultados obtidos com 7 configurações específicas de conjuntos de classificadores, combinando as duas bases de treinamento utilizadas e diversos modos de representação do texto.

Tabela 6.11: Resultados do *Ensemble* de Classificação da Dimensão *Extraversion*

Conjunto	**Acurácia**	**Classificador**	**Modo**	**Base**
Ext1	55,56%	MLPC	LIWC	ESSAYS
Ext2	53,33%	GNB	*Word2Vec*	ESSAYS
Ext3	53,33%	kNN	*unigram*	ESSAYS
		LR	*unigram*	ESSAYS
Ext4	57,78%	LR	LIWC	ESSAYS
		GNB	LIWC	ESSAYS
		kNN	LIWC	ESSAYS
Ext5	60,00%	LR	LIWC	ESSAYS
		GNB	LIWC	ESSAYS
		kNN	LIWC	ESSAYS
		SVM	*unigram*	ESSAYS
Ext6	**68,89%**	LR	LIWC	ESSAYS
		GNB	LIWC	ESSAYS
		kNN	LIWC	ESSAYS
		SVM	*unigram*	ESSAYS
		MLPC	*Word2Vec*	ESSAYS
Ext7	51,11%	LR	LIWC	*myPersonality*

Para a dimensão *Extraversion*, os melhores resultados foram obtidos com a combinação das três técnicas de representação com a base ESSAYS sendo utilizada no treinamento dos classificadores.

O conjunto "Ext6" apresentou uma acurácia de 69%, sendo que nos trabalhos investigados, o melhor resultado para esta dimensão também foi observado no trabalho apresentado por Tandera et al. (2017), que obteve uma acurácia de 65% utilizando redes neurais com a base *myPersonality* e representação LIWC.

6.8.4 Resultados obtidos com a dimensão *Agreeableness*

A Tabela 6.12 apresenta os resultados obtidos com 10 configurações específicas de conjuntos de classificadores, combinando as duas bases de treinamento utilizadas e diversos modos de representação do texto.

Tabela 6.12: Resultados do *Ensemble* de Classificação da Dimensão *Agreeableness*

Conjunto	**Acurácia**	**Classificador**	**Modo**	**Base**
Agr1	57,78%	LR	LIWC	ESSAYS
Agr2	60,00%	GNB	LIWC	ESSAYS
Agr3	62,22%	GNB	*bigram*	ESSAYS
Agr4	55,56%	GNB	*Word2Vec*	ESSAYS
Agr5	51,11%	kNN	LIWC	ESSAYS
		LR	LIWC	ESSAYS
		MLPC	LIWC	ESSAYS
Agr6	60,00%	GNB	LIWC	ESSAYS
		GNB	*bigram*	ESSAYS
		GNB	*Word2Vec*	ESSAYS
Agr7	**66,67%**	kNN	LIWC	ESSAYS
		LR	LIWC	ESSAYS
		RFOREST	LIWC	ESSAYS
		MLPC	*Word2Vec*	ESSAYS
Agr8	35,56%	kNN	LIWC	*myPersonality*
		GNB	LIWC	*myPersonality*
		MLPC	LIWC	*myPersonality*
Agr9	55,56%	kNN	LIWC	*myPersonality*
		GNB	LIWC	*myPersonality*
		MLPC	LIWC	*myPersonality*
		LR	*unigram*	*myPersonality*
Agr10	51,11%	kNN	LIWC	*myPersonality*
		GNB	LIWC	*myPersonality*
		MLPC	LIWC	*myPersonality*
		LR	*unigram*	*myPersonality*
		GNB	*Word2Vec*	*myPersonality*

Para esta dimensão, os resultados obtidos com a base ESSAYS também apresentaram os melhores resultados na identificação. No caso da representação, a combinação de LIWC e Word2Vec, apresentado como "Agr7", ocasionou o melhor resultado, apresentando uma acurácia de 67%. Nos trabalhos investigados, o melhor resultado para esta dimensão foi de 65%, também obtido no trabalho de Tandera et al. (2017) utilizando redes neurais, base *myPersonality* e representação LIWC.

6.8.5 Resultados obtidos com a dimensão *Neuroticism*

A Tabela 6.13 apresenta os resultados obtidos com 11 configurações específicas de conjuntos de classificadores, combinando as duas bases de treinamento utilizadas e diversos modos de representação do texto.

Tabela 6.13: Resultados do *Ensemble* de Classificação da Dimensão *Neuroticism*

Conjunto	**Acurácia**	**Classificador**	**Modo**	**Base**
Neu1	57,78%	MLPC	LIWC	ESSAYS
Neu2	48,89%	GNB	LIWC	ESSAYS
Neu3	42,22%	kNN	LIWC	ESSAYS
Neu4	62,22%	kNN	*unigram*	ESSAYS
Neu5	60,00%	GNB	*Word2Vec*	ESSAYS
Neu6	**68,89%**	MLPC	LIWC	ESSAYS
		kNN	*unigram*	ESSAYS
		LR	*unigram*	ESSAYS
		MLPC	*unigram*	ESSAYS
Neu7	53,33%	GNB	*bigram*	ESSAYS
		LR	*bigram*	ESSAYS
		LR	*trigram*	ESSAYS
		GNB	LIWC	ESSAYS
		GNB	*Word2Vec*	ESSAYS
Neu8	60,00%	GNB	LIWC	ESSAYS
		GNB	*bigram*	ESSAYS
		GNB	*Word2Vec*	ESSAYS
Neu9	57,78%	kNN	LIWC	*myPersonality*
Neu10	57,78%	LR	LIWC	*myPersonality*
Neu11	51,11%	kNN	LIWC	*myPersonality*
		GNB	LIWC	*myPersonality*
		MLPC	LIWC	*myPersonality*
		LR	*unigram*	*myPersonality*
		GNB	*Word2Vec*	*myPersonality*

No caso do *Neuroticism*, os melhores resultados também foram obtidos com a base ESSAYS, sendo que neste caso o conjunto “Neu6”, que agregou a representação LIWC com *unigram*, apresentou o melhor resultado, com uma acurácia de 69%. Como referência, o melhor resultado observado na literatura investigada foi novamente verificado no trabalho de Tandera et al. (2017), com um valor de 65%, utilizando redes neurais, base *myPersonality* e representação LIWC.

6.9 Publicações

O trabalho descrito nesta tese, originou artigos aceitos em conferências científicas de abrangência internacional. O primeiro deles, aceito e apresentado no XXVIII Simpósio Brasileiro de Informática na Educação (SBIE 2017), intitulado "Identificação de estilo de aprendizagem: Um modelo de inferência automatizado baseado no perfil de personalidade identificado nos textos produzidos pelo aluno" (Buiar et al., 2017), apresenta a primeira versão do modelo de identificação de personalidade proposto na presente pesquisa, denominado IPP, sendo aplicado na realização da inferência do Estilo de Aprendizagem, de acordo com o modelo de Felder e Silverman (Felder et al., 1988).

O segundo trabalho, apresenta a utilização da identificação da personalidade, utilizando o modelo IPP inicialmente proposto, para a identificação das dimensões *Extraversion* e *Openness* que são utilizadas como referência para o sequenciamento adaptativo dos objetos de aprendizagem, em colaboração com a pesquisadora Zenaide Silva. Foi aceito e apresentado na XXII Conferência Internacional sobre Informática na Educação (TISE 2017) com o título "Adaptação da Interface de Objetos de Aprendizagem a partir do Perfil de Personalidade do Aprendiz" (Silva et al., 2017).

O terceiro trabalho a ser publicado, "Detecção automática de traços de personalidade e recomendação de agrupamento com o modelo Big Five", foi aceito para apresentação XXIX Simpósio Brasileiro de Informática na Educação (SBIE 2018). Realizado em conjunto com a pesquisadora Taís Ferreira (UFU), aborda os resultados dos experimentos realizados pelos autores, na identificação de personalidade a partir do texto para a formação de grupos de colaboração. Apresentando os resultados obtidos na presente pesquisa, este trabalho foi selecionado entre os seis artigos que irão concorrer ao melhor artigo deste simpósio.

6.10 Considerações

O experimento realizado com a base UNIVERSIDADE teve como objetivo verificar o comportamento do modelo IP3 na realização da identificação do perfil de personalidade de alunos, em um ambiente real. Para isto foram realizados testes de identificação com diversas combinações de classificadores, bases de treinamento e formas de representação do texto, sendo que somente os resultados de alguns conjuntos relevantes e representativos das diversas combinações obtidas, foram apresentados nas tabelas apresentadas na Seção 6.8.

A natureza dos textos dos alunos, obtidos a partir das atividades educacionais registradas no *Moodle*, correspondendo à base UNIVERSIDADE, tem uma característica diferenciada em relação às duas bases de referência utilizadas para treinamento dos classificadores. A base ESSAYS foi formada com redações realizadas pelos voluntários, em que estes referiam-se a si próprios, utilizando uma linguagem formal. No caso da base *myPersonality*, os textos que originaram esta base, foram obtidos do ambiente *Facebook*, onde foram verificados textos em formato mais livre, sem uma preocupação dos autores com formalismos sintáticos.

Por sua vez, os textos da base UNIVERSIDADE são oriundos de atividades educacionais, onde os alunos muitas vezes realizavam a descrição de características de terceiros, o que motivou a utilização das duas bases distintas de classificação e verificação de qual base seria mais adequada. Foi verificado que em todas as dimensões do BIG FIVE, a utilização da base ESSAYS apresentou melhores resultados na identificação da personalidade dos alunos, a partir dos textos das atividades educacionais, tendo como referência a base utilizada neste experimento.

Em relação aos classificadores, foi verificado que não houve destaque em relação a um classificador específico, nem mesmo grandes melhorias nos resultados com mudança nos

parâmetros dos classificadores. Mas a utilização de conjuntos de classificadores, utilizando formas de representação diferenciadas e combinadas, permitiu a obtenção de melhores resultados no processo de identificação. Apesar dos resultados inferiores obtidos com a representação de forma isolada das técnicas nGRAM e *Word2Vec*, em relação a representação LIWC, a combinação destas técnicas possibilitou a melhoria dos resultados obtidos, comparado com os valores obtidos utilizando somente o LIWC. A Tabela 6.14 apresenta os valores observados nas pesquisas investigadas, que utilizaram as bases ESSAYS e *myPersonality*, comparados com os resultados obtidos na presente pesquisa.

Tabela 6.14: Comparativo da Acurácia Obtida nos Experimentos Investigados

	OPN	**CON**	**EXT**	**AGR**	**NEU**
Mairesse et al. (2007)	63%	56%	56%	56%	58%
Alam et al. (2013)	69%	59%	59%	58%	63%
Iacobelli e Culotta (2013)	62%	56%	61%	53%	56%
Tighe et al. (2016)	61%	55%	54%	57%	57%
Majumder et al. (2017)	63%	57%	59%	57%	59%
Tandera et al. (2017)	74%	56%	65%	59%	65%
Yu e Markov (2017)	71%	51%	61%	54%	61%
Ensemble IP3	**91%**	**62%**	**68%**	**66%**	**68%**

Os valores obtidos com o processo de identificação do perfil de personalidade, utilizando o modelo BIG FIVE, dos alunos presentes na base UNIVERSIDADE, por meio do modelo IP3, tendo como referência de treinamento dos classificadores, as bases ESSAYS e *myPersonality*, demonstra resultados superiores de acurácia, se comparados com as demais pesquisas identificadas, que realizaram experimentos com as bases ESSAYS e *myPersonality*. Nesta comparação também deve ser levado em consideração que a presente pesquisa utilizou a base UNIVERSIDADE para a obtenção da acurácia, ao passo que os demais experimentos citados utilizaram a separação das bases ESSAYS e *myPersonality* em conjuntos de treinamento e teste. Além disto, nos testes para validação do modelo IP3, foi comprovada a viabilidade da utilização de uma base de treinamento em idioma inglês, no processo de classificação de uma base em idioma português, utilizando as técnicas utilizadas pelo modelo desenvolvido.

7 Considerações Finais

A escolha do tema de pesquisa, encontrando-se na interseção entre as áreas de conhecimento da Computação, da Educação e da Psicologia, exige que o ambiente de delimitação da pesquisa seja bem definido, considerando que estas áreas têm contextos extremamente abrangentes. A conjunção da personalidade com os ambientes de educação a distância e os sistemas computacionais criam um tripé, que por si só, abrem espaço para amplas e profundas análises. Acredita-se que este é apenas um ponto inicial de uma longa curva de um processo de aprendizagem e desenvolvimento deste novo paradigma, a modelagem educacional por meio de sistemas computacionais, para personificar o processo de ensino-aprendizagem onde o foco, mais importante não são os instrumentos educacionais em si, mas o processo individual de aprendizagem do estudante. O estado da arte parece nos mostrar que apenas estamos nos primeiros passos de um processo revolucionário e transformador, para um melhor sistema de ensino aprendizado, saindo do modelo clássico de ensinar por meio de uma via, onde o centro é o professor, para um modelo completamente individualizado, onde o que mais importa é como a personalidade do aluno desenvolve o modelo pessoal de aprendizado, podendo ser potencializado pela utilização de sistemas computacionais.

Nesta pesquisa foi verificado que a identificação do perfil de personalidade dos alunos é um importante recurso a ser utilizado nos processos educacionais, possuindo relações com o desempenho acadêmico, estilos de aprendizado, sistemas adaptativos, recomendação e retenção acadêmica. Também foi constatado que a pouca disponibilidade de bases de referência consistentes, em português, tem sido um fator inibidor da utilização em maior escala da identificação automática do perfil de personalidade nas instituições brasileiras. Foi constatado que as principais iniciativas de identificação do perfil de personalidade, de forma automática, utilizaram bases de referência obtidas a partir de informações obtidas em redes sociais ou de redações em ambientes controlados.

A presente pesquisa foi desenvolvida tendo como base a hipótese de viabilidade da identificação automática do perfil de personalidade de alunos, baseada nos textos por estes redigidas em atividades educacionais. Inicialmente, foi realizado um estudo sobre os alicerces teóricos que envolvem a personalidade e os seus relacionamentos com a computação e a educação, bem como das técnicas e processos que permitem esta identificação. Esta fundamentação colaborou para o desenvolvimento das bases iniciais do modelo apresentado.

A partir da revisão da literatura abrangendo a relação entre a Personalidade e a Educação, foi constatada a relevância da realização da identificação da personalidade dos alunos em ambientes educacionais. Uma segunda revisão da literatura, voltada à Identificação da Personalidade a partir de Texto, permitiu verificar quais os modelos de personalidade, bases de dados, idiomas e técnicas de classificação foram adotados nas pesquisas investigadas. Os resultados destes levantamentos permitiram a evolução do modelo proposto.

Foi então realizada a especificação de um modelo computacional, denominado IP3, com o objetivo de realizar a identificação automática do perfil de personalidade dos alunos. Este modelo foi idealizado com o objetivo de permitir uma flexibilidade na escolha de uma ou mais bases de treinamento, de diversos idiomas, bem como, propiciar a escolha de diversas

formas de representação do texto e técnicas de classificação. De uma forma parametrizável, o modelo permite a combinação das bases de treinamento, formas de representação e técnicas de classificação que sejam mais adequadas ao cenário em que for utilizado.

Durante a pesquisa foram avaliados os modelos de representação de texto para a aplicação de técnicas de aprendizado de máquina, abrangendo a utilização do léxico e métodos estatísticos baseados em processamento de linguagem natural, as quais foram utilizadas no modelo apresentado. Também foram investigadas as estratégias de classificação mais adequadas para a aplicação de aprendizado de máquina no texto representado, sendo proposto o conceito *ensemble* de classificadores para o modelo IP3.

A pesquisa realizada sobre bases de referência, evidenciou a dificuldade em obter uma base consistente em português, adequada ao processo de identificação dos textos dos alunos, o que direcionou esta pesquisa para a utilização de bases em outros idiomas. Foram selecionadas duas bases em inglês para utilização nos experimentos realizados, comprovando a viabilidade da utilização de bases de treinamento diversas. A especificação do modelo IP3 considerou que a forma de representação do texto deveria permitir a utilização de idiomas distintos entre a base de treinamento e o conjunto de textos a ser identificado.

Os testes iniciais realizados, somente com as bases de referência ESSAYS e *myPersonality*, permitiram comprovar que os resultados apresentados pelo modelo estavam compatíveis com as pesquisas verificadas no estado da arte, propiciando a evolução da pesquisa no sentido de verificar a identificação com textos educacionais em um ambiente real. Para isto, foi conduzido um experimento inicial em uma turma semi-presencial, evoluindo na sequencia para um estudo mais abrangente em três turmas presenciais que utilizavam apoio do ambiente virtual na condução das atividades propostas pelos professores. Com a aplicação do inventário manual, foram obtidos os valores correspondentes ao perfil de personalidade destes alunos de acordo com o BIG FIVE. Estes valores foram utilizados como referência para validação do modelo IP3 quando da identificação do perfil de personalidade a partir dos textos obtido nas atividades educacionais. Utilizando conjuntos de classificadores e formas de representação distintas, para cada uma das dimensões do modelo BIG FIVE, os melhores resultados foram obtidos com a base de treinamento ESSAYS.

No estudo de caso realizado foi comprovada a viabilidade da utilização do modelo IP3 neste ambiente educacional de forma automática e não intrusiva, sendo que os resultados obtidos foram superiores aos valores encontrados na literatura, em experimentos de identificação de personalidade a partir do texto, em ambientes não educacionais.

Desta forma, os objetivos desejados com a presente pesquisa foram obtidos, apresentando resultados relevantes para a comunidade científica. Nas pesquisas realizadas não foram identificadas iniciativas que comprovassem a identificação automática do perfil de personalidade de alunos, em idioma português, baseado somente nos registros dos textos obtidos a partir das atividades educacionais. Acredita-se que esta iniciativa inédita poderá contribuir de forma prática nos avanços da área da computação da personalidade aplicada ao ensino.

7.1 Trabalhos Futuros

A partir da presente pesquisa, vislumbra-se diversos trabalhos futuros que podem ser realizados. Em primeiro lugar, uma ampliação do experimento poderia ser realizado, considerando cursos de diversas áreas como público alvo, para verificar a distribuição dos perfis de personalidade dos alunos em função das áreas específicas dos diversos cursos.

O desenvolvimento do modelo proposto na forma de um *plugin* para o ambiente *Moodle* poderia permitir a utilização da identificação da personalidade dos alunos diretamente no AVA,

propiciando uma ferramenta que auxilie aos professores e tutores na utilização dos indicadores de personalidade na condução das atividades educacionais.

A utilização do modelo proposto poderia ser utilizada para a obtenção de uma base mais consistente do que as utilizadas, e em idioma português, para disponibilizar de forma pública uma base de referência em idioma português para a condução de outras pesquisas na área de identificação de personalidade.

A aplicação do modelo proposto na formação de grupos de colaboração, a partir da identificação do perfil individual baseado em texto, e a avaliação dos resultados obtidos pelos grupos em atividades educacionais poderia ser utilizado para investigar o impacto da personalidade na formação destes grupos.

A identificação de personalidade de forma não invasiva, com base nos textos dos alunos em um ambiente virtual de aprendizagem poderia ser utilizado para obter indicadores sobre retenção associada a personalidade, propiciando ações preventivas contra a evasão elevada que ocorre nestes ambientes.

Referências

Afonso, S., Bick, E., Haber, R. e Santos, D. (2002). Floresta sintá (c) tica: a treebank for portuguese. Em *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)(Las Palmas de Gran Canaria Espanha 29-31 de Maio de 2002)*. ELRA.

Aguiar, J. J. B. (2017). Considerando estilos de aprendizagem, emoções e personalidade em informática na educação. *Informática na educação: teoria & prática*, 20(2).

Al-Dujaily, A. e Ryu, H. (2008). A study on personality in designing adaptive e-learning systems. Em *Advanced Learning Technologies, 2008. ICALT'08. Eighth IEEE International Conference on*, páginas 136–138. IEEE.

Alam, F., Stepanov, E. A. e Riccardi, G. (2013). Personality traits recognition on social network-facebook. *WCPR (ICWSM-13), Cambridge, MA, USA*.

Allen, I. E. e Seaman, C. A. (2007). Likert scales and data analyses. *Quality progress*, 40(7):64.

Allport, G. W. (1937). *Personality*. Holt New York.

Altanopoulou, P. e Tselios, N. (2015). How does personality affect wiki-mediated learning? Em *Interactive Mobile Communication Technologies and Learning (IMCL), 2015 International Conference on*, páginas 16–18. IEEE.

Andrade, J. M. d. (2008). *Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil*. Tese de doutorado, Universidade de Brasília.

Antunes, C. (1998). *Inteligências Múltiplas E Seus Estímulos (as)*. Papirus Editora.

Argamon, S., Dhawle, S., Koppel, M. e Pennebaker, J. W. (2005). Lexical predictors of personality type. Em *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*.

Balage Filho, P. P., Pardo, T. A. S. e Aluísio, S. M. (2013). An Evaluation of the Brazilian portuguese LIWC Dictionary for Sentiment Analysis. *9th Brazilian Synposium in Information and Human Language Technology*, páginas 215–219.

Bartol, A. M. e Bartol, C. R. (2014). *Criminal behavior: A psychological approach*. Boston: Pearson, c2014. xxiii, 644 pages: illustrations; 24 cm.

Bird, S., Klein, E. e Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York.

Blickle, G. (1996). Personality traits, learning stratigies, and performance. *European Journal of personality*, 10(5):337–352.

Boyd, R. (2014). Meh: Meaning extraction helper (version 1.0. 6)[software].

Boyle, G. J., Matthews, G. e Saklofske, D. H. (2008). *The Sage handbook of personality theory and assessment: Personality measurement and testing*, volume 2. Sage.

Bradley, M. M. e Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. Relatório técnico, Citeseer.

Brody, N. (2000). History of theories and measurements of intelligence. *Handbook of intelligence*, páginas 16–33.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Univ of California Press.

Buiar, J., Pimentel, A. e Oliveira, L. (2017). Identificação de estilo de aprendizagem: Um modelo de inferência automatizado baseado no perfil de personalidade identificado nos textos produzidos pelo aluno. Em *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, página 1157.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B. e Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. Em *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, páginas 108–122.

Busato, V. V., Prins, F. J., Elshout, J. J. e Hamaker, C. (1998). The relation between learning styles, the Big Five personality traits and achievement motivation in higher education. *Personality and Individual Differences*, 26(1):129–140.

Carro, R. M. e Sanchez-Horreo, V. (2017). The effect of personality and learning styles on individual and collaborative learning: Obtaining criteria for adaptation. Em *Global Engineering Education Conference (EDUCON), 2017 IEEE*, páginas 1585–1590. IEEE.

Cassidy, S. (2004). Learning styles: An overview of theories, models, and measures. *Educational psychology*, 24(4):419–444.

Cattell, R. B. (1957). *Personality and motivation structure and measurement.* World Book Co.

Cattell, R. B., Eber, H. W. e Tatsuoka, M. (1970). Handbook for the 16 personality factor questionnaire. *Champaign, IL: Institute for Personality and Ability Testing*.

Cavnar, W. B., Trenkle, J. M. et al. (1994). N-gram-based text categorization. *Ann arbor mi*, 48113(2):161–175.

Cerqueira, T. C. S. (2000). *Estilos de aprendizagem em universitários*. Tese de doutorado, Universidade Estadual de Campinas, Faculdade de Educação.

Chamorro-Premuzic, T. e Furhnam, A. (2006). Intellectual competence and the intelligent personality: A third way in differential psychology. *Review of General Psychology*, 10(3):251.

Chamorro-Premuzic, T. e Furnham, A. (2003). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality*, 37(4):319 – 338.

Chen, G., Davis, D., Hauff, C. e Houben, G.-J. (2016). On the impact of personality in massive open online learning. Em *Proceedings of the 2016 conference on user modeling adaptation and personalization*, páginas 121–130. ACM.

Chung, C. K. e Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of research in personality*, 42(1):96–132.

Coffield, F., Moseley, D., Hall, E., Ecclestone, K. et al. (2004). Learning styles and pedagogy in post-16 learning: A systematic and critical review.

Cohen, A. e Baruth, O. (2017). Personality, learning, and satisfaction in fully online academic courses. *Computers in Human Behavior*, 72:1–12.

Conaway, M. R. (1990). A random effects model for binary data. *Biometrics*, páginas 317–328.

Cortes, C. e Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Costa, P. e McCrae, R. (1989). Neo five-factor inventory (neo-ffi). *Odessa, FL: Psychological Assessment Resources*.

Costa, P. T. e McCrae, R. R. (1985). *The NEO personality inventory*. PAR Psychological Assessment Resources.

Costa, P. T. e McCrae, R. R. (1992). Neo personality inventory–revised (neo-pi-r) and neo five-factor inventory (neo-ffi) professional manual. *Odessa, FL: Psychological Assessment Resources*.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, páginas 215–242.

Crozier, W. R. (2013). *Individual learners: Personality differences in education*. Routledge.

De Raad, B. (2000). *The Big Five Personality Factors: The psycholexical approach to personality.* Hogrefe & Huber Publishers.

De Raad, B. e Schouwenburg, H. C. (1996). Personality in learning and education: A review. *European Journal of personality*, 10(5):303–336.

Deary, I. J. (2009). The trait approach to personality. *The Cambridge handbook of personality psychology*, páginas 89–109.

Delahaye, B. e Thompson, B. (1991). Learning styles?what do they measure? *Asia Pacific Journal of Human Resources*, 29(2):60–68.

Di Giunta, L., Alessandri, G., Gerbino, M., Kanacri, P. L., Zuffiano, A. e Caprara, G. V. (2013). The determinants of scholastic achievement: The contribution of personality traits, self-esteem, and academic self-efficacy. *Learning and individual Differences*, 27:102–108.

Digman, J. M. (2002). *Historical antecedents of the five-factor model.* American Psychological Association.

Donche, V., Maeyer, S., Coertjens, L., Daal, T. e Petegem, P. (2013). Differential use of learning strategies in first-year higher education: The impact of personality, academic motivation, and teaching strategies. *British Journal of Educational Psychology*, 83(2):238–251.

Du, J., Shi, R., Zhen, Y. e Lu, W. (2017). An analysis of influence factors for academic performance about personality traits and thinking styles of students: Use a c programming language course in college as an example. Em *Computer Science and Education (ICCSE), 2017 12th International Conference on*, páginas 167–171. IEEE.

Dunn, R., Beaudry, J. S. e Klavas, A. (2002). Survey of research on learning styles. *California Journal of Science Education*, 2(2):75–98.

Dunn, R., Dunn, K. e Price, G. E. (1977). Diagnosing learning styles: A prescription for avoiding malpractice suits. *The Phi Delta Kappan*, 58(5):418–420.

Dusay, J. M. (1972). Egograms and the "constancy hypothesis". *Transactional Analysis Bulletin*, 2(3):37–41.

Ekman, P., Friesen, W. V., O'sullivan, M. e Scherer, K. (1980). Relative importance of face, body, and speech in judgments of personality and affect. *Journal of personality and social psychology*, 38(2):270.

Esuli, A. e Sebastiani, F. (2007). Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation*, páginas 1–26.

Eysenck, H. (1970). The structure of human personality methuen. *London, UK*.

Eysenck, H. (1978). The development of personality and its relation to learning. *Critical Studies in Education*, 20(1):134–181.

Eysenck, H. J. (1947). *Dimensions of personality: A record of research carried out in collaboration with ht himmelweit [and others]*. K. Paul, Trench, Trubner.

Eysenck, H. J. e Eysenck, S. B. (1976). *Eysenck personality questionnaire*. Educational and industrial testing service.

Eysenck, S. B., Eysenck, H. J. e Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and individual differences*, 6(1):21–29.

Ezpeleta, E., Zurutuza, U. e Hidalgo, J. M. G. (2016). Short messages spam filtering using personality recognition. Em *Proceedings of the 4th Spanish Conference on Information Retrieval*, página 7. ACM.

Farias, A. B., Dobrões, J. A. L. e da Silva, R. Y. F. (2013). Strategies for teaching based on academic personality types. Em *XVIII Conferência Internacional sobre Informática na Educação (TISE)*, páginas 633–636.

Farnadi, G., Zoghbi, S., Moens, M.-F. e De Cock, M. (2013). Recognising personality traits using facebook status updates. Em *Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13)*. AAAI.

Felder, R. M., Silverman, L. K. et al. (1988). Learning and teaching styles in engineering education. *Engineering education*, 78(7):674–681.

Felder, R. M. e Soloman, B. (2006). Index of learning styles. 1991.

Ferreira, T. e Fernandes, M. (2017). Detecção de traços de personalidade em textos para apoiar a formação de grupos para colaboração. Em *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, página 1627.

Floudas, C. A. e Visweswaran, V. (1995). Quadratic optimization. Em *Handbook of global optimization*, páginas 217–269. Springer.

Funder, D. C. (1997). *The personality puzzle.* WW Norton & Co.

Furnham, A., Monsen, J. e Ahmetoglu, G. (2009). Typical intellectual engagement, Big Five personality traits, approaches to learning and cognitive ability predictors of academic performance. *British Journal of Educational Psychology*, 79(4):769–782.

Gagniuc, P. A. (2017). *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons.

Galton, F. (1949). *The Measurement of Character.* Prentice-Hall, Inc.

Gardner, H. e Hatch, T. (1989). Educational implications of the theory of multiple intelligences. *Educational researcher*, 18(8):4–10.

Garner, S. (1995). Weka: The waikato environment for knowledge analysis. Em *Proceedings of the New Zealand computer science research students conference*, páginas 57–64. Citeseer.

Golbeck, J. (2016). Predicting personality from social media text. *AIS Transactions on Replication Research*, 2(1):2.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, 2(1):141–165.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R. e Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96.

Gosling, S. D., Rentfrow, P. J. e Swann Jr., W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504–528.

Guilford, J. P. (1975). Factors and factors of personality. *Psychological Bulletin*, 82(5):802.

Hall, C. W., Kauffmann, P. J., Wuensch, K. L., Swart, W. E., DeUrquidi, K. A., Griffin, O. H. e Duncan, C. S. (2015). Aptitude and personality traits in retention of engineering students. *Journal of Engineering Education*, 104(2):167–188.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. e Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hayes, J. e Allinson, C. W. (1988). Cultural differences in the learning styles of managers. *Management International Review*, páginas 75–80.

He, H. e Garcia, E. A. (2008). Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, 21(9):1263–1284.

Heaven, P. C., Ciarrochi, J. e Vialle, W. (2007). Conscientiousness and eysenckian psychoticism as predictors of school grades: A one-year longitudinal study. *Personality and Individual Differences*, 42(3):535–546.

Hockenbury, D. H. e Hockenbury, S. E. (2010). *Discovering psychology*. Macmillan.

Hoekstra, H., Ormel, J. e De Fruyt, F. (1996). Neo persoonlijkheidsvragenlijsten neo-pi-r en neo-ffi. handleiding. Em *Lisse: Swets & Zeitlinger*.

Iacobelli, F. e Culotta, A. (2013). Too neurotic, not too friendly: structured personality classification on textual data. Em *Proc of Workshop on Computational Personality Recognition, AAAI Press, Melon Park, CA*, páginas 19–22.

Isbister, K. e Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International journal of human-computer studies*, 53(2):251–267.

John, O. P., Angleitner, A. e Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European journal of Personality*, 2(3):171–203.

John, O. P. e Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.

Joulin, A., Grave, E., Bojanowski, P. e Mikolov, T. (2016). Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759.

Jung, C. G. (2014). *Collected Works of CG Jung, Volume 17: Development of Personality: Development of Personality*, volume 17. Princeton University Press.

Kågström, J., Karlsson, R. e Kågström, E. (2013). uClassify web service.

Kalghatgi, M. P., Ramannavar, M. e Sidnal, N. S. (2015). A neural network approach to personality prediction based on the big-five model. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, 2(8):56–63.

Kappe, R. e van der Flier, H. (2010). Using multiple and specific criteria to assess the predictive validity of the big five personality factors on academic performance. *Journal of Research in Personality*, 44(1):142–145.

Karsvall, A. (2002). Personality preferences in graphical interface design. Em *Proceedings of the second Nordic conference on Human-computer interaction*, páginas 217–218. ACM.

Kassin, S. M. (2003). *Essentials of psychology*. Prentice Hall.

Keerthi, S., Shevade, S., Bhattacharyya, C. e Murthy, K. (2001). Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649.

Keirsey, D. e Bates, M. M. (1984). *Please understand me: Character & temperament types*. Prometheus Nemesis Book Company Del Mar, CA.

Kim, J., Lee, A. e Ryu, H. (2013). Personality and its effects on learning performance: Design guidelines for an adaptive e-learning system based on a user model. *International Journal of Industrial Ergonomics*, 43(5):450–461.

Kirschner, P. A. (2017). Stop propagating the learning styles myth. *Computers & Education*, 106:166–171.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Em *Ijcai*, volume 14, páginas 1137–1145. Montreal, Canada.

Komarraju, M. e Karau, S. J. (2005). The relationship between the big five personality traits and academic motivation. *Personality and Individual Differences*, 39(3):557–567.

Komisin, M. C. e Guinn, C. I. (2012). Identifying personality types using document classification methods. Em *FLAIRS Conference*.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V. e Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543.

Kristensen, C. H., Gomes, C. F. d. A., Justo, A. R. e Vieira, K. (2011). Brazilian norms for the affective norms for english words. *Trends in Psychiatry and Psychotherapy*, 33(3):135–146.

Laidra, K., Pullmann, H. e Allik, J. (2007). Personality and intelligence as predictors of academic achievement: A cross-sectional study from elementary to secondary school. *Personality and individual differences*, 42(3):441–451.

Lima, A. e Castro, L. (2014). A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural Networks*, 58:122–130.

Lior, R. et al. (2007). *Data Mining With Decision Trees: Theory And Applications*, volume 69. World Scientific.

Liu, F., Perez, J. e Nowson, S. (2016). A language-independent and compositional model for personality trait recognition from short texts. *arXiv preprint arXiv:1610.04345*.

Luyckx, K. e Daelemans, W. (2008). Using syntactic features to predict author personality from text. *Proceedings of Digital Humanities*, 2008:146–9.

Magalhães, E., Portela, M., Oliveira, P., Salgueira, A. e Costa, M. (2009). Personalidade, género e desempenho académico: um estudo com estudantes de medicina portugueses. Em *X Congresso Internacional Galego-Português de Psicopedagogia*.

Magnini, B. e Cavaglia, G. (2000). Integrating subject field codes into wordnet. Em *LREC*, páginas 1413–1418.

Mairesse, F. e Walker, M. (2006). Words mark the nerds: Computational models of personality recognition through language. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, páginas 543–548.

Mairesse, F., Walker, M. A., Mehl, M. R. e Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, páginas 457–500.

Majumder, N., Poria, S., Gelbukh, A. e Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79.

Matthews, G., Deary, I. J. e Whiteman, M. C. (2003). *Personality traits*. Cambridge University Press.

McAdams, D. P. e Olson, B. D. (2010). Personality development: Continuity and change over the life course. *Annual review of psychology*, 61:517–542.

McAdams, D. P. e Pals, J. L. (2006). A new big five: fundamental principles for an integrative science of personality. *American psychologist*, 61(3):204.

McCrae, R. R. e Costa, P. T. (2003). *Personality in adulthood: A five-factor theory perspective*. Guilford Press.

McCrae, R. R. e Terracciano, A. (2005). Universal features of personality traits from the observer's perspective: data from 50 cultures. *Journal of personality and social psychology*, 88(3):547.

McCulloch, W. S. e Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Melo, S. d., Dantas, A. C. e Fernandes, M. (2017). Modelo do estudante baseado em emoções e perfis de personalidade para recomendação de estratégias pedagógicas personalizadas. Em *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, página 967.

Messick, S. (1984). The nature of cognitive styles: Problems and promise in educational practice. *Educational Psychologist*, 19(2):59–74.

Meyer, D. e Wien, F. T. (2017). Support vector machines.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W. e Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, 56(2):128.

Michel, W., Shoda, Y. e Smith, R. (2004). Introduction to personality: Toward an integration.

Mikolov, T., Chen, K., Corrado, G. e Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, A. (1991). Personality types, learning styles and educational goals. *Educational Psychology*, 11(3-4):217–238.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Minamikawa, A. e Yokoyama, H. (2011). Blog tells what kind of personality you have: egogram estimation from japanese weblog. Em *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, páginas 217–220. ACM.

Mischel, W. (2013). *Personality and assessment*. Psychology Press.

Moffitt, K., Giboney, J., Ehrhardt, E., Burgoon, J. e Nunamaker, J. (2012). Structured programming for linguistic cue extraction (splice). Em *Proceedings of the HICSS-45 Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium*, páginas 103–108.

Munezero, M. D., Montero, C. S., Sutinen, E. e Pajunen, J. (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111.

Murray, H. A. (1938). *Explorations in personality.* Oxford Univ. Press.

Myers, I. B., McCaulley, M. H. e Most, R. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*, volume 1985. Consulting Psychologists Press Palo Alto, CA.

Myers, L. e Sirois, M. J. (2006). Spearman correlation coefficients, differences between. *Wiley StatsRef: Statistics Reference Online*.

Nasrabadi, N. M. (2007). Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4):049901.

Nass, C. e Lee, K. M. (2001). Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171–181.

Nijboer, F., Morin, F. O., Carmien, S. P., Koene, R. A., Leon, E. e Hoffmann, U. (2009). Affective brain-computer interfaces: Psychophysiological markers of emotion in healthy persons and in persons with amyotrophic lateral sclerosis. Em *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, páginas 1–11. IEEE.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, 66(6):574.

Nov, O., Arazy, O., López, C. e Brusilovsky, P. (2013). Exploring personality-targeted ui design in online social participation systems. Em *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, páginas 361–370. ACM.

Nunes, M. A. S. N., Teles, F. R. e de Souza, J. G. (2013). Inferindo personalidade via tweets. *Revista GEINTEC-Gestão, Inovação e Tecnologias*, 3(3):045–057.

Oberlander, J. e Nowson, S. (2006). Whose Thumb is It Anyway?: Classifying Author Personality from Weblog Text. Em *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, páginas 627–634, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ortony, A., Clore, G. L. e Foss, M. A. (1987). The referential structure of the affective lexicon. *Cognitive science*, 11(3):341–364.

Ostendorf, F. e Angleitner, A. (1994). Psychometric properties of the german translation of the neo personality inventory (neo-pi-r). Em *Poster presented at the Seventh Conference of the European Association for Personality Psychology, Madrid, Spain*.

O'Connor, M. C. e Paunonen, S. V. (2007). Big five personality predictors of post-secondary academic performance. *Personality and Individual differences*, 43(5):971–990.

Pang, B., Lee, L. et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Pasqualotti, P. R. e Vieira, R. (2008). Wordnetaffectbr: uma base lexical de palavras de emoções para a língua portuguesa. *RENOTE*, 6(1).

Patterson, J. e Gibson, A. (2017). *Deep Learning: A Practitioner's Approach*. O'Reilly Media, Inc.

Peabody, D. (1970). Evaluative and descriptive aspects in personality perception: a reappraisal. *Journal of personality and social psychology*, 16(4):639.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. e Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennebaker, J., Booth, R., Boyd, R. e Francis, M. (2015). Linguistic inquiry and word count: Liwc2015. Relatório técnico, Pennebaker Conglomerates, Austin, TX, USA.

Pennebaker, J. W., Booth, R. J. e Francis, M. E. (2007). Linguistic inquiry and word count: Liwc [computer software]. Relatório técnico, Pennebaker Conglomerates, Austin, TX,USA.

Pennebaker, J. W., Francis, M. E. e Booth, R. J. (1993). Linguistic inquiry and word count : Liwc. Relatório técnico, Southern Methodist University, Dallas, TX, USA.

Pennebaker, J. W., Francis, M. E. e Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Pennebaker, J. W. e King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312.

Perrenoud, P. (2001). *A pedagogia na escola das diferenças: fragmentos de uma sociologia do fracasso*. Artmed.

Petrov, S., Das, D. e McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. Em Schoelkopf, B., Burges, C. e Smola, A., editores, *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Poria, S., Gelbukh, A., Agarwal, B., Cambria, E. e Howard, N. (2013). Common sense knowledge based personality recognition from text. Em *Mexican International Conference on Artificial Intelligence*, páginas 484–496. Springer.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological bulletin*, 135(2):322.

Porto, S. M., Costa, W. S., Nunes, M. e Matos, L. N. (2011). Como a extração de personalidade através do teclado pode beneficiar a personalização na educação. *Anais do XXII SBIE-XVII WIE*.

Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234.

Raento, M., Oulasvirta, A. e Eagle, N. (2009). Smartphones: An emerging tool for social scientists. *Sociological methods & research*, 37(3):426–454.

Rammstedt, B. e John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212.

Rangel, F., Rosso, P., Potthast, M., Stein, B. e Daelemans, W. (2015). Overview of the 3rd author profiling task at pan 2015. Em *CLEF*, página 2015. sn.

Reeves, B. e Nass, C. (1998). *The Media equation: how people treat computers, television, and new media*. Cambridge University Press.

Refaeilzadeh, P., Tang, L. e Liu, H. (2009). Cross-validation. Em *Encyclopedia of database systems*, páginas 532–538. Springer.

Řehůřek, R. e Sojka, P. (2011). Gensim—statistical semantics in python.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Saadé, R. G., Kira, D., Nebebe, F. e Otrakji, C. (2006). Openness to experience: An hci experiment. *Issues in informing science & information technology*, 3.

Saez, Y., Navarro, C., Mochon, A. e Isasi, P. (2014). A system for personality and happiness detection. *IJIMAI*, 2(5):7–15.

Salleh, N., Mendes, E., Grundy, J. e Burch, G. S. J. (2010). An empirical study of the effects of conscientiousness in pair programming using the five-factor personality model. Em *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*, páginas 577–586. ACM.

Salton, G., Wong, A. e Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Saucier, G. e Goldberg, L. R. (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of personality*, 69(6):847–879.

Saucier, G. e Srivastava, S. (2015). What makes a good structural model of personality? evaluating the big five and alternatives. *Handbook of personality and social psychology*, 3:283–305.

Schmitt, D. P., Allik, J., McCrae, R. R. e Benet-Martínez, V. (2007). The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of cross-cultural psychology*, 38(2):173–212.

Schultz, D. P. e Schultz, S. E. (2016). *Theories of personality*. Cengage Learning.

Shami, N. S., Hancock, J. T., Peter, C., Muller, M. e Mandryk, R. (2008). Measuring affect in hci: going beyond the individual. Em *CHI'08 extended abstracts on Human factors in computing systems*, páginas 3901–3904. ACM.

Silva, Z., Buiar, J., Pimentel, A. e Ferreira, L. (2017). Adaptação de objetos de aprendizagem a partir do perfil de personalidade do aprendiz. Em *XXII Conferência Internacional sobre Informática na Educação*, volume 13, páginas 367–372. Nuevas Ideas en Informática Educativa.

Silva, Z. C. (2017). *Adaptação de apresentação de conteúdos de objeto de aprendizagem considerando estilos de aprendizagem*. Tese de doutorado, Tese de doutorado em computação. Universidade Federal do Paraná (UFPR). Ciência da Computação (UFPR).

Skinner, B. F. (1982). Sobre o behaviorismo, tradução de maria da penha villalobos.

Tandera, T., Suhartono, D., Wongso, R., Prasetio, Y. L. et al. (2017). Personality prediction system from facebook users. *Procedia Computer Science*, 116:604–611.

Tausczik, Y. R. e Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Taylor, A. e MacDonald, D. A. (1999). Religion and the five factor model of personality: An exploratory investigation using a canadian university sample. *Personality and Individual Differences*, 27(6):1243–1259.

Thorndike, E. L. (2013). *The elements of psychology*, volume 126. Routledge.

Thurstone, L. (1948). Primary mental abilities. *Science (New York, NY)*, 108(2813):585.

Tighe, E. P., Ureta, J. C., Pollo, B. A. L., Cheng, C. K. e de Dios Bulos, R. (2016). Personality trait classification of essays with the application of feature reduction. Em *SAAIP@ IJCAI*, páginas 22–28.

Tomlinson, M. T., Hinote, D. e Bracewell, D. B. (2013). Predicting conscientiousness through semantic analysis of facebook posts. *Proceedings of WCPR*.

Travaglia, L. C. (2016). Composição tipológica de textos como atividade de formulação textual. *Revista do GELNE*, 4(1):1–10.

Uleman, J. S., Adil Saribay, S. e Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annu. Rev. Psychol.*, 59:329–360.

Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

Uspensky, J. V. (1937). *Introduction to mathematical probability*. McGraw-Hill Book Company, New York.

Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C. e Vallieres, E. F. (1992). The academic motivation scale: A measure of intrinsic, extrinsic, and amotivation in education. *Educational and psychological measurement*, 52(4):1003–1017.

Van Rossum, G. e Drake, F. L. (2003). *Python language reference manual*. Network Theory.

Varela, O. E., Cater, J. J. e Michel, N. (2012). Online learning in management education: an empirical study of the role of personality traits. *Journal of Computing in Higher Education*, 24(3):209–225.

Verhoeven, B., Daelemans, W. e De Smedt, T. (2013). Ensemble methods for personality recognition. Em *Proceedings of the workshop on computational personality recognition*, páginas 35–38.

Vinciarelli, A. e Mohammadi, G. (2014). A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Vonderwell, S. e Zachariah, S. (2005). Factors that influence participation in online learning. *Journal of Research on Technology in education*, 38(2):213–230.

Weber, M. R. (2015). The relationship between personality and student learning. *Journal of Hospitality & Tourism Education*, 27(4):135–146.

Wei, H., Zhang, F., Yuan, N. J., Cao, C., Fu, H., Xie, X., Rui, Y. e Ma, W.-Y. (2017). Beyond the words: Predicting user personality from heterogeneous information. Em *Proceedings of the tenth ACM international conference on web search and data mining*, páginas 305–314. ACM.

Wilson, M. (1988). Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, 20(1):6–10.

Witten, I. H., Frank, E., Hall, M. A. e Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wolf, M. B. e Ackerman, P. L. (2005). Extraversion and intelligence: A meta-analytic investigation. *Personality and Individual Differences*, 39(3):531–542.

Wylie, A. (2014). How to make your copy more readable: Make sentences shorter.

Yu, J. e Markov, K. (2017). Deep learning based personality recognition from facebook status updates. Em *The 8th International Conference on Awareness Science and Technology*.

Zhang, L. (2003). Does the big five predict learning approaches? *Personality and Individual Differences*, 34(8):1431–1446.

Apêndice A: Formulário BFI-44

Este apêndice apresenta o formulário utilizado no levantamento do Perfil de Personalidade dos alunos. Este formulário foi baseado no inventário BFI-44 (John e Srivastava, 1999), de avaliação da personalidade, utilizando o modelo BIG FIVE. A adaptação deste inventário para o idioma português é valido para o contexto brasileiro, de acordo com o trabalho realizado Andrade (2008), que utilizou um universo de 5.089 respondentes das cinco regiões brasileiras para condução da pesquisa.

Nome: _____

INSTRUÇÕES DE PREENCHIMENTO. Inicialmente preencha o seu nome no campo acima. Depois verifique que a seguir encontram-se algumas características que podem, ou não, lhe dizer respeito. Por favor, escolha um dos números (de 1 a 5), da escala abaixo, que **melhor expresse** a sua verdadeira personalidade e anote-o no espaço ao lado de cada afirmação. Vale ressaltar que não existem respostas certas ou erradas, mas sim, respostas que mais se identificam com você mesmo. Utiliza a seguinte escala de resposta:

1	2	3	4	5
Discordo totalmente	Discordo em parte	Nem concordo nem discordo	Concordo em parte	Concordo totalmente

Eu me vejo como alguém que ...

___ 01 é conversador, comunicativo.
___ 02 às vezes é frio e distante.
___ 03 tende a ser crítico com os outros.
___ 04 é minucioso, detalhista no trabalho.
___ 05 é assertivo, não teme expressar o que sente.
___ 06 insiste até concluir a tarefa ou o trabalho.
___ 07 é depressivo, triste.
___ 08 gosta de cooperar com os outros.
___ 09 é original, tem sempre novas ideias.
___ 10 é temperamental, muda de humor facilmente.
___ 11 é inventivo, criativo.
___ 12 é reservado.
___ 13 valoriza o artístico, o estético.
___ 14 é emocionalmente estável, não se altera facilmente.
___ 15 é prestativo e ajuda os outros.
___ 16 às vezes, é tímido, inibido.
___ 17 pode ser um tanto descuidado.
___ 18 é amável, tem consideração pelos outros.
___ 19 tende a ser preguiçoso.
___ 20 faz as coisas com eficiência.
___ 21 é relaxado, controla bem o estresse.
___ 22 é facilmente distraído.
___ 23 mantém-se calmo nas situações tensas.
___ 24 prefere trabalho rotineiro.
___ 25 é curioso sobre muitas coisas diferentes.
___ 26 é sociável, extrovertido.
___ 27 é geralmente confiante.
___ 28 às vezes, é rude (grosseiro) com os outros.
___ 29 é cheio de energia.
___ 30 começa discussões, disputas com os outros.
___ 31 é um trabalhador de confiança.
___ 32 faz planos e os segue à risca.
___ 33 tem uma imaginação fértil.
___ 34 fica tenso com frequência.
___ 35 é engenhoso, alguém que gosta de analisar profundamente as coisas.
___ 36 fica nervoso facilmente.
___ 37 gera muito entusiasmo.
___ 38 tende a ser desorganizado.
___ 39 gosta de refletir, brincar com as ideias.
___ 40 tem capacidade de perdoar, perdoa fácil.
___ 41 preocupa-se muito com tudo.
___ 42 tende a ser quieto, calado.
___ 43 tem poucos interesses artísticos.
___ 44 é sofisticado em artes, música ou literatura.

Figura A.1: Formulário BFI-44 Adaptado

Apêndice B: Resultados nGRAM

Este apêndice apresenta o detalhamento dos resultados obtidos com o experimento descrito na Seção 6.6.

Tabela B.1: Base ESSAYS com *unigram*

	Openness $\overline{acc}$ (%)	***Conscientiousness*** $\overline{acc}$ (%)	***Extraversion*** $\overline{acc}$ (%)	***Agreeableness*** $\overline{acc}$ (%)	***Neuroticism*** $\overline{acc}$ (%)
BASELINE	51,2±1,8	50,1±2,2	51,0±2,5	51,4±0,7	49,0±0,9
kNN	48,5±2,0	50,4±2,2	48,3±1,6	51,3±0,4	**51,8**±2,3
GNB	49,5±1,7	49,4±1,8	49,7±1,5	52,1±1,9	49,6±2,9
LR	51,5±1,2	49,2±0,7	50,1±1,3	52,4±0,6	49,5±1,8
RFOREST	50,4±2,2	48,8±1,8	50,6±2,5	51,4±2,0	51,1±1,2
MLPC	**51,5**±0,0	**50,8**±0,0	**51,7**±0,0	**53,1**±0,1	50,1±0,2
SVM	**51,5**±0,0	**50,8**±0,0	**51,7**±0,0	**53,1**±0,1	49,7±0,6

Tabela B.2: Base ESSAYS com *bigram*

	Openness $\overline{acc}$ (%)	***Conscientiousness*** $\overline{acc}$ (%)	***Extraversion*** $\overline{acc}$ (%)	***Agreeableness*** $\overline{acc}$ (%)	***Neuroticism*** $\overline{acc}$ (%)
BASELINE	51,2±1,8	50,1±2,2	51,0±2,5	51,4±0,7	49,0±0,9
kNN	49,7±3,5	48,7±2,1	48,6±2,0	48,1±2,0	**52,1**±3,1
GNB	49,5±1,2	48,9±1,1	49,3±1,2	52,1±1,2	49,7±1,4
LR	**51,5**±2,5	48,4±1,2	50,1±0,5	52,0±1,1	50,1±2,0
RFOREST	50,9±1,6	**52,5**±1,1	50,1±2,9	51,9±0,3	49,5±3,5
MLPC	47,5±0,8	49,1±1,7	47,6±1,5	50,0±1,9	51,4±2,6
SVM	51,5±0,0	50,8±0,0	**51,7**±0,0	**53,1**±0,1	49,7±0,5

Tabela B.3: Base ESSAYS com *trigram*

	Openness $\overline{acc}$ (%)	***Conscientiousness*** $\overline{acc}$ (%)	***Extraversion*** $\overline{acc}$ (%)	***Agreeableness*** $\overline{acc}$ (%)	***Neuroticism*** $\overline{acc}$ (%)
BASELINE	51,2±1,8	50,1±2,2	51,0±2,5	51,4±0,7	49,0±0,9
kNN	51,1±1,8	49,0±0,7	50,5±1,7	49,5±2,4	49,4±3,5
GNB	50,4±1,1	49,5±1,7	50,5±1,3	50,4±3,1	50,7±1,6
LR	50,8±2,2	49,2±0,8	49,4±0,5	52,2±1,0	50,0±2,0
RFOREST	50,5±1,7	49,2±1,7	50,7±0,9	52,6±1,4	50,4±2,1
MLPC	49,8±2,0	50,1±2,3	48,0±1,9	51,7±1,2	**51,2**±1,8
SVM	**51,5**±0,0	**50,8**±0,0	**51,7**±0,0	**53,1**±0,1	49,6±0,8

Tabela B.4: Base *myPersonality* com *unigram*

	Openness $\overline{acc}$ (%)	***Conscientiousness*** $\overline{acc}$ (%)	***Extraversion*** $\overline{acc}$ (%)	***Agreeableness*** $\overline{acc}$ (%)	***Neuroticism*** $\overline{acc}$ (%)
BASELINE	50,4±0,7	49,5±0,5	49,9±0,9	49,9±0,4	49,6±0,3
kNN	66,4±0,9	49,9±0,7	51,7±1,3	49,9±0,7	54,3±1,6
GNB	73,5±1,0	53,0±1,3	55,7±2,1	52,7±1,6	60,9±1,3
LR	**74,3**±0,0	53,6±1,0	57,3±0,6	52,6±1,5	**62,5**±0,0
RFOREST	**74,3**±0,0	53,7±0,3	57,4±0,4	**53,1**±1,1	62,5±0,1
MLPC	74,3±0,0	52,8±0,8	**57,5**±0,0	52,5±1,3	**62,5**±0,0
SVM	**74,3**±0,0	**54,1**±0,0	**57,5**±0,0	53,0±0,3	**62,5**±0,0

Tabela B.5: Base *myPersonality* com *bigram*

	Openness $\overline{acc}$ (%)	***Conscientiousness*** $\overline{acc}$ (%)	***Extraversion*** $\overline{acc}$ (%)	***Agreeableness*** $\overline{acc}$ (%)	***Neuroticism*** $\overline{acc}$ (%)
BASELINE	50,4±0,7	49,5±0,5	49,9±0,9	49,9±0,4	49,6±0,3
kNN	66,2±1,2	51,4±1,7	52,2±0,4	49,4±1,7	54,7±1,1
GNB	57,6±12,5	53,5±0,5	52,6±2,6	49,8±2,4	40,5±0,9
LR	74,3±0,1	52,3±1,3	55,9±0,9	51,2±1,3	62,0±0,3
RFOREST	**74,3**±0,0	53,6±0,4	**57,6**±0,1	52,9±0,3	**62,5**±0,0
MLPC	73,5±0,3	51,1±2,3	54,3±0,9	50,8±0,8	60,9±0,9
SVM	**74,3**±0,0	**54,1**±0,0	57,5±0,0	**53,1**±0,0	**62,5**±0,0

Tabela B.6: Base *myPersonality* com *trigram*

	Openness $\overline{acc}$ (%)	***Conscientiousness*** $\overline{acc}$ (%)	***Extraversion*** $\overline{acc}$ (%)	***Agreeableness*** $\overline{acc}$ (%)	***Neuroticism*** $\overline{acc}$ (%)
BASELINE	50,4±0,7	49,5±0,5	49,9±0,9	49,9±0,4	49,6±0,3
kNN	67,8±0,9	51,0±0,9	52,0±1,6	50,3±1,4	54,3±1,1
GNB	30,0±0,6	47,2±0,5	43,9±0,9	48,7±0,9	40,7±0,8
LR	**74,3**±0,2	52,4±1,2	55,8±1,0	51,6±0,9	61,3±0,1
RFOREST	74,3±0,0	**54,2**±0,3	57,5±0,1	**53,3**±0,2	**62,5**±0,0
MLPC	74,3±0,0	50,6±0,7	52,8±1,3	53,1±0,0	54,5±1,5
SVM	74,3±0,0	54,1±0,0	**57,5**±0,0	53,1±0,0	**62,5**±0,0

Apêndice C: Resultados *Word2Vec*

Este apêndice apresenta o detalhamento dos resultados obtidos com o experimento descrito na Seção 6.7.

Tabela C.1: Comparativo da Base ESSAYS com *Word2Vec*

Tamanho	**10**	**20**	**50**	**100**	**200**	**300**	**400**	**500**
	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)
Openness								
BASELINE	51,2±1,8	51,2±1,8	51,2±1,8	51,2±1,8	51,2±1,8	51,2±1,8	51,2±1,8	51,2±1,8
kNN	48,2±1,8	52,4±1,6	50,1±2,1	50,1±2,1	51,4±1,6	52,3±2,8	50,5±2,2	51,8±2,4
GNB	52,3±3,5	52,3±2,2	51,0±2,8	52,1±2,7	51,5±2,4	51,2±2,7	51,3±2,2	51,2±2,9
LR	**54,8**±2,0	52,8±1,2	51,6±1,5	53,9±2,3	54,0±1,6	54,0±1,8	53,0±1,5	54,4±2,3
RFOREST	50,8±2,1	50,7±3,3	50,9±2,6	52,5±2,2	50,1±1,3	51,4±1,9	52,1±2,7	51,3±1,6
MLPC	50,0±2,2	51,9±2,0	51,4±1,1	52,1±1,1	51,5±0,0	51,5±0,0	51,5±0,0	52,9±1,3
SVM	53,0±2,4	52,1±1,3	51,6±1,6	51,7±2,3	51,0±1,5	51,0±0,9	51,0±1,0	51,2±0,4
Conscientiousness								
BASELINE	50,1±2,2	50,1±2,2	50,1±2,2	50,1±2,2	50,1±2,2	50,1±2,2	50,1±2,2	50,1±2,2
kNN	50,1±2,6	49,1±1,8	48,7±2,1	48,4±0,8	49,0±1,8	50,9±1,4	51,3±2,4	49,9±1,4
GNB	48,7±1,4	51,2±1,2	50,1±1,9	49,6±1,1	48,6±0,6	48,4±1,4	49,6±1,6	50,0±1,3
LR	50,4±1,6	50,5±1,5	50,3±2,1	52,7±2,2	50,9±2,8	51,2±2,7	50,6±1,6	**53,7**±2,0
RFOREST	50,4±2,1	49,3±1,8	49,7±1,5	51,5±2,8	48,9±1,9	48,7±0,9	52,0±2,3	50,3±1,0
MLPC	50,0±1,3	50,7±2,9	48,9±1,9	50,2±2,3	50,8±0,0	50,8±0,0	50,8±0,0	52,5±0,8
SVM	51,2±2,4	52,2±2,3	49,9±1,0	50,0±0,8	50,5±0,6	50,8±0,0	50,8±0,0	50,8±0,0
Extraversion								
BASELINE	51,0±2,5	51,0±2,5	51,0±2,5	51,0±2,5	51,0±2,5	51,0±2,5	51,0±2,5	51,0±2,5
kNN	49,7±0,7	48,8±1,5	48,6±2,4	50,1±1,3	48,2±1,5	49,9±1,0	48,8±2,0	48,9±2,2
GNB	51,2±2,2	51,7±2,1	50,1±1,2	50,1±1,9	50,0±1,8	50,8±1,5	50,7±1,4	51,2±2,2
LR	51,3±1,5	50,6±1,7	49,6±1,2	52,3±1,5	50,8±2,3	49,9±2,8	47,6±1,3	51,0±1,6
RFOREST	49,3±0,6	50,9±3,5	50,3±3,6	52,3±1,6	**52,5**±2,7	51,2±1,3	51,4±0,8	49,3±2,4
MLPC	49,7±1,8	50,2±1,4	50,5±2,5	51,2±1,5	51,7±0,0	51,7±0,0	51,7±0,0	49,3±2,9
SVM	51,0±2,0	50,5±0,8	50,8±1,5	51,3±0,5	51,5±0,3	51,8±0,1	51,7±0,1	51,7±0,1
Agreeableness								
BASELINE	51,4±0,7	51,4±0,7	51,4±0,7	51,4±0,7	51,4±0,7	51,4±0,7	51,4±0,7	51,4±0,7
kNN	51,0±2,2	50,8±2,5	50,2±3,1	47,5±1,7	49,7±1,8	50,0±2,1	50,3±2,3	50,1±1,8
GNB	51,0±1,1	49,7±2,4	50,8±1,9	51,1±2,5	52,1±2,1	50,3±1,8	49,8±2,5	51,3±1,9
LR	52,1±1,3	52,7±1,4	51,4±2,1	51,8±1,6	52,4±1,0	52,0±2,3	51,8±1,3	51,8±0,9
RFOREST	51,1±2,4	50,9±1,8	52,3±0,6	51,3±1,2	51,5±0,8	52,0±1,1	52,5±1,5	50,7±1,0
MLPC	51,4±2,5	51,8±1,8	51,3±2,5	50,0±1,8	**53,1**±0,1	**53,1**±0,1	**53,1**±0,1	51,9±2,2
SVM	52,8±1,1	52,9±0,6	53,0±0,3	52,9±0,3	53,1±0,1	**53,1**±0,1	**53,1**±0,1	**53,1**±0,1
Neuroticism								
BASELINE	49,0±0,9	49,0±0,9	49,0±0,9	49,0±0,9	49,0±0,9	49,0±0,9	49,0±0,9	49,0±0,9
kNN	51,4±2,2	51,2±1,4	52,2±1,6	49,8±2,2	50,1±2,3	50,3±2,5	48,8±2,6	49,8±3,1
GNB	51,0±2,9	51,6±2,0	51,5±1,9	51,0±1,8	51,4±1,8	52,0±1,9	51,6±1,7	51,2±1,5
LR	52,8±1,7	51,1±0,9	49,9±1,1	52,2±2,2	51,5±0,7	**54,4**±2,8	50,3±2,5	51,3±2,1
RFOREST	49,3±1,7	51,8±1,3	50,8±1,5	52,4±1,5	49,7±1,7	51,0±2,3	49,5±0,9	51,3±1,9
MLPC	49,4±1,0	48,8±1,9	50,2±1,2	52,0±1,5	50,0±0,0	50,0±0,0	50,0±0,0	51,8±2,2
SVM	53,0±2,0	51,2±1,4	51,1±1,5	52,3±1,9	51,3±2,2	51,9±1,7	51,4±2,2	51,8±2,0

Tabela C.2: Comparativo da Base *myPersonality* com *Word2Vec*

Tamanho	**10**	**20**	**50**	**100**	**200**	**300**	**400**	**500**
	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)	$\overline{acc}$ (%)
Openness								
BASELINE	50,4±0,5	50,4±0,5	50,4±0,5	50,4±0,5	50,4±0,5	50,4±0,5	50,4±0,5	50,4±0,5
kNN	66,7±0,7	66,9±0,3	66,8±0,5	66,1±0,5	66,6±0,6	66,3±0,7	66,7±0,5	66,9±1,0
GNB	73,6±0,5	71,0±2,1	66,6±1,3	60,8±6,2	57,8±7,2	56,8±7,6	56,3±8,0	55,8±7,9
LR	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0
RFOREST	74,3±0,0	**74,3**±0,0	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0
MLPC	74,3±0,0	74,1±0,2	74,2±0,2	74,2±0,2	74,1±0,2	73,7±0,4	74,1±0,4	74,1±0,3
SVM	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0	74,3±0,0
Conscientiousness								
BASELINE	49,6±0,5	49,6±0,5	49,6±0,5	49,6±0,5	49,6±0,5	49,6±0,5	49,6±0,5	49,6±0,5
kNN	51,7±1,0	49,9±1,4	50,1±1,0	51,4±0,8	50,1±1,1	50,9±0,8	51,4±1,1	50,5±0,5
GNB	49,9±2,4	48,8±2,4	48,8±2,0	48,6±2,2	48,8±2,2	48,8±2,3	48,5±2,1	48,5±2,1
LR	53,3±0,4	53,5±0,5	53,5±0,5	53,5±0,5	53,7±0,3	53,6±0,4	53,5±0,4	53,6±0,3
RFOREST	53,3±0,5	53,3±0,6	53,2±0,8	53,3±0,7	53,3±0,5	53,2±0,5	53,2±0,9	53,6±0,6
MLPC	53,8±0,5	52,1±1,6	52,0±1,0	52,7±1,1	**54,1**±0,0	53,9±0,4	53,6±0,6	52,5±1,0
SVM	53,4±0,6	53,7±0,4	53,9±0,1	54,0±0,1	54,0±0,1	54,0±0,1	54,0±0,1	54,0±0,0
Extraversion								
BASELINE	49,7±0,9	49,7±0,9	49,7±0,9	49,7±0,9	49,7±0,9	49,7±0,9	49,7±0,9	49,7±0,9
kNN	52,2±0,5	51,8±0,8	51,6±1,5	52,2±0,5	51,6±1,2	50,9±1,2	51,8±1,1	52,1±0,6
GNB	49,4±2,4	47,0±2,2	46,2±2,1	46,4±1,6	46,4±1,5	46,0±1,4	46,0±1,4	46,0±1,7
LR	57,2±0,7	57,0±0,7	57,1±0,7	57,0±0,8	57,2±0,5	57,1±0,7	57,0±0,8	57,0±0,8
RFOREST	57,1±0,7	57,3±0,3	57,3±0,5	57,1±0,7	57,1±0,7	57,5±0,5	57,4±0,4	57,1±0,6
MLPC	57,4±0,4	55,8±1,0	56,2±1,2	56,1±1,4	**57,6**±0,0	57,6±0,2	56,8±0,9	56,4±1,6
SVM	57,2±1,0	57,4±0,3	**57,6**±0,0	**57,6**±0,0	**57,6**±0,0	**57,6**±0,0	**57,6**±0,0	**57,6**±0,0
Agreeableness								
BASELINE	49,8±0,5	49,8±0,5	49,8±0,5	49,8±0,5	49,8±0,5	49,8±0,5	49,8±0,5	49,8±0,5
kNN	49,9±0,7	50,6±1,3	50,1±1,3	50,0±0,7	50,4±0,5	50,5±0,9	49,5±1,1	50,4±0,9
GNB	52,0±2,4	52,0±1,9	51,2±2,0	51,0±2,3	51,1±2,6	51,4±2,6	51,4±2,3	51,4±2,6
LR	52,3±0,9	52,2±1,2	52,1±1,1	52,4±1,3	52,4±1,5	52,4±1,3	52,2±1,1	52,1±1,2
RFOREST	52,0±1,1	52,1±1,0	52,4±1,3	52,5±1,2	52,6±0,8	52,5±1,1	52,3±1,1	52,1±1,0
MLPC	52,1±2,1	50,9±1,0	50,4±1,0	51,9±1,4	52,1±1,7	53,0±0,9	53,1±0,0	52,5±1,7
SVM	52,7±0,9	53,0±0,8	53,2±0,4	53,3±0,2	**53,3**±0,2	53,2±0,1	53,2±0,1	53,2±0,1
Neuroticism								
BASELINE	49,8±0,4	49,8±0,4	49,8±0,4	49,8±0,4	49,8±0,4	49,8±0,4	49,8±0,4	49,8±0,4
kNN	54,9±0,9	55,0±1,1	55,6±0,6	54,8±1,5	54,9±1,5	54,7±1,5	54,0±0,8	54,4±0,8
GNB	61,3±1,0	60,1±1,4	58,8±1,2	57,2±2,0	56,2±2,4	55,7±2,2	55,7±2,3	55,8±2,2
LR	62,3±0,5	62,3±0,5	62,3±0,5	62,3±0,4	62,4±0,4	62,3±0,5	62,3±0,5	62,3±0,5
RFOREST	62,4±0,2	62,2±0,5	62,4±0,3	62,4±0,2	62,3±0,3	62,3±0,4	62,4±0,2	62,4±0,2
MLPC	62,1±0,7	61,6±1,0	62,1±0,6	61,7±0,6	**62,5**±0,0	62,4±0,1	61,3±0,9	61,4±1,1
SVM	62,3±0,3	62,4±0,4	62,5±0,1	62,5±0,1	62,5±0,1	62,5±0,1	62,5±0,1	62,5±0,1